

Master's Programme in Computer, Communication and Information Sciences

Evaluating privacy and ethical implications of AI-assisted documentation in healthcare

A Case Study

Leevi Pulkkinen

Copyright © 2025 Leevi Pulkkinen

Author Leevi Pulkkinen

Title Evaluating privacy and ethical implications of AI-assisted documentation in healthcare — A Case Study

Degree programme Computer, Communication and Information Sciences

Major Computer Science

Supervisor Prof. Tom Bäckström

Advisor MEng Matts Kotkamaa

Collaborative partner Vitec Raisoft Oy

Date 12 December 2025 **Number of pages** 50+4 **Language** English

Abstract

Clinical documentation is an essential part of patient care, but can be time consuming and reduce interaction time with patients. With the emergence of artificial intelligence (AI) technologies, such as automatic speech recognition (ASR), natural language processing (NLP), and large language models (LLMs), promising new tools have been proposed to make clinical documentation more efficient and patient centric. However, the introduction of AI powered documentation tools in healthcare settings raises a multitude of concerns regarding data protection, privacy, and ethics. This thesis examines these implications by conducting a Data Ethics Decision Aid (DEDA) evaluation of an AI assisted documentation tool called Smart Care and contrasting the findings with the overlying literature. The findings emphasized the importance of lawful data processing and robustness of AI services against attacks as key factors to maintaining the privacy of individuals. In addition, human oversight, accountability, and proper training were highlighted to ensure responsible use. Finally, transparency towards patients and proper acquisition of consent were seen as essential for creating trust between patients, users, and technology. Further research in practical settings is needed to perform a sophisticated evaluation of these issues and the proposed solutions.

Keywords Clinical documentation, artificial intelligence, privacy, ethics

Tekijä Leevi Pulkkinen

Työn nimi Tekoälyavusteisen dokumentoinnin yksityisyyden ja eettisten vaikutusten arviointi terveydenhuollossa — Tapaustutkimus

Koulutusohjelma Computer, Communication and Information Sciences

Pääaine Computer Science

Työn valvoja Prof. Tom Bäckström

Työn ohjaaja YAMK Matts Kotkamaa

Yhteistyötaho Vitec Raisoft Oy

Päivämäärä 12.12.2025

Sivumäärä 50+4

Kieli englanti

Tiivistelmä

Kliininen dokumentointi on olennainen osa potilashoitoa, mutta se voi olla aikaa vievää ja vähentää vuorovaikutusaikaa potilaiden kanssa. Tekoälyteknologioiden, kuten automaattisen puheentunnistuksen (ASR), luonnollisen kielen käsittelyn (NLP) ja suurten kielimallien (LLM), esiintulon myötä on ehdotettu lupaavia uusia työkaluja, jotka voisivat tehdä kliinisestä dokumentoinnista tehokkaampaa ja potilaskeskeisempää. Tekoälypohjaisten dokumentointityökalujen käyttöönotto terveydenhuoltoympäristössä herättää kuitenkin useita huolenaiheita tietoturvan, yksityisyyden ja etiikan suhteen. Tässä opinnäytetyössä tarkastellaan näitä seurauksia tekemällä Data Ethics Decision Aid (DEDA) arviointi tekoälyavusteisesta dokumentointityökalusta nimeltä Smart Care ja vertaamalla tuloksia taustalla olevaan kirjallisuuteen. Tulokset korostivat, että lainmukainen tiedonkäsittely ja hyökkäyksiltä suojattu tekoälypalvelu ovat keskeisiä tekijöitä yksilöiden yksityisyyden turvaamisessa. Lisäksi ihmiskontrollia, vastuullisuutta ja asianmukaista kouluttamista pidettiin tärkeinä tekijöinä vastuullisen käytön varmistamiseksi. Lopuksi läpinäkyvyyttä potilaita kohtaan ja suostumuksen asianmukaista hankkimista pidettiin olennaisina luottamuksen luomisessa potilaiden, käyttäjien ja teknologian välille. Lisätutkimusta käytännön kontekstissa tarvitaan näiden ongelmien ja mahdollisten ratkaisujen syvällistä arviointia varten.

Avainsanat Kliininen dokumentointi, tekoäly, yksityisyys, etiikka

Preface

Firstly, I would like to thank my thesis supervisor Tom Bäckström for his patient and supportive guidance throughout this long process. Secondly, I want to thank Vitec Raisoft for providing an engaging project and the opportunity to turn it into a thesis topic. Thank you to all colleagues and partners who contributed to the project.

I also want to thank my family and friends for their encouragement and for keeping me moving forward. A special thanks to our dog Hiili for her sharp-eyed supervision of my work and for making sure that I got some fresh air on our morning walks.

Otaniemi, 12 December 2025

Leevi Pulkkinen

Contents

Abstract	3
Abstract (in Finnish)	4
Preface	5
Contents	6
1 Introduction	8
I. Literature Review	9
2 AI-assisted documentation in healthcare	9
2.1 Speech recognition	9
2.2 AI scribes	11
2.3 Large language models	12
3 Implications of AI-assisted documentation for privacy and ethics	14
3.1 Data protection, privacy and robustness	14
3.1.1 General Data Protection Regulation	14
3.1.2 Security, robustness and privacy preserving techniques	18
3.1.3 Consent	20
3.2 Responsible AI use in healthcare documentation	21
3.2.1 Transparency, explainability, and trust	21
3.2.2 Accountability and the role of AI in healthcare documentation	23
3.2.3 Fairness	25
4 Data Ethics Decision Aid (DEDA) framework	26
II. Case Study: Smart Care prototype evaluation	28

5	Methods	28
6	The Smart Care prototype	30
6.1	Project background and stakeholders	30
6.2	Goals	31
6.3	System requirements	31
6.4	AI service integration	32
6.5	Implementation results	33
7	DEDA assessment results	35
7.1	Project overview and values	35
7.2	Data-related considerations	36
7.3	General considerations	39
7.4	DEDA assessment conclusion	41
8	Discussion	42
9	Conclusion	46
	References	47
A	DEDA assessment questions	51

1 Introduction

Clinical documentation is a central part of quality care, as it enables the recording of care procedures, ensures continuity, enhances patient safety, and supports clinical decision-making. Documentation tasks, such as data entry, reviewing notes, and billing, are typically performed using electronic health record (EHR) systems [1]. Physicians have been shown to spend nearly twice as much time on EHR and desk work compared to direct patient care [2]. Documentation-related activities consume a significant share of healthcare professionals' time and may reduce interaction with patients. With the emergence of artificial intelligence (AI) technologies, such as automatic speech recognition (ASR), natural language processing (NLP), and large language models (LLMs), promising new tools have been proposed to make clinical documentation more efficient and patient centric.

However, the use of AI technologies for healthcare documentation raises a multitude of concerns about privacy, data protection, and ethics. AI algorithms are dependent on high-quality training data [3]. As healthcare data is highly sensitive in nature, training AI systems using healthcare data creates a considerable risk to the privacy of individuals [3]. In addition, the implementation of AI-powered services in healthcare is frequently managed by external commercial entities, which have an increased role in accessing, processing, and safeguarding patient data [4]. Possible data leakage or misuse can cause significant damage to patients, healthcare providers, and software vendors [3].

Trustworthiness is also an important consideration when AI systems are used in healthcare. Large language models (LLMs) pose the risk of generating outputs that are not based on any factual information and regardless of correctness, they are presented with high confidence [5]. In addition, LLMs can contain bias due to biased training data [5]. Furthermore, many AI systems are "black boxes" by nature, and the reasoning for their outputs can be unclear to users. The black box model reduces trust in AI systems [6] and makes it difficult to obtain informed consent from patients, since it is challenging for patients and clinicians to understand the full extent behind the rationales of AI systems [7].

The objective of this thesis is to evaluate the implications of using AI-powered tools in healthcare documentation, with a focus on privacy, data protection, and ethics. The benefits and limitations of existing AI technologies are explored based on recent literature. In addition, a case study is conducted that involves the development of the **Smart Care** prototype, which integrates speech-to-text functionality and AI-based text structuring into an existing healthcare documentation system. The ethical and privacy implications of this implementation are assessed using the Data Ethics Decision Aid (DEDA) assessment [8]. Finally, the findings from the literature and the case study are compared and discussed to outline key considerations for the responsible use of AI in healthcare documentation.

I. Literature Review

2 AI-assisted documentation in healthcare

Data entry has been shown to be the most time consuming activity related EHR use [1]. Family medicine professionals have been found to spend almost half of their working day using an EHR system [9]. In a study by Arndt et al. [9], documentation was listed as the most time consuming activity (of 15 different categories) at 23.7% of total EHR use time. Another study showed that physicians spent nearly twice the time on EHR and desk work compared to direct clinical face time with patients [2]. Even during face-to-face consultations, 37% of the physician's time was allocated to EHR and desk work [2]. The impact of EHR systems on the workflows of individuals is highly dependent on their role, clinical setting, and the usability of the EHR system [1]. Although individual experiences vary, tasks related to documentation and data entry constitute a significant workload for healthcare professionals. This workload and the emergence of artificial intelligence (AI) have prompted interest in AI tools designed to make clinical documentation more efficient and higher quality, allowing for better and more focused care.

2.1 Speech recognition

Many healthcare facilities have adopted speech recognition (SR) software as an alternative to traditional documentation methods, such as typing and medical scribes, in hopes of reducing costs and time spent on EHR documentation [10]. In medical contexts, transcription typically refers to the generation of electronic medical records and reports by converting a doctor's speech or doctor-patient interactions into text [11]. SR is used to transcribe speech into text in real time, and it offers promising opportunities for generating medical documentation through speech [12]. The proposed benefits of using SR with EHR systems include improved accuracy, recording at the point-of-care, standardization of documentation, and lower costs of patient services [12]. In this section, the use, benefits, and drawbacks of speech recognition technologies for healthcare documentation are explored.

For decades, speech technology has been developed as a subfield of signal processing, aiming to enable machines to recognize and analyze human speech [13]. Typically, speech technology systems include three major components: pre-processing, feature extraction, and machine learning (ML) algorithms. Pre-processing involves techniques such as noise suppression, silence removal, and channel equalization, which enhance the robustness and efficiency of speech-based systems. Feature extraction converts speech signals into features, typically divided into linguistic features (for example, words and their grammatical alterations) and acoustic features (for example, melody, tone, and jitter). Finally, ML models are used to map these features into outputs.

Historically, hidden Markov model (HMM) and Gaussian mixture model (GMM) based models were most commonly used, but today deep learning (DL) methods are an essential component of automatic speech recognition (ASR) and other speech processing tasks. [13]

The main advantage of SR transcription technology is reduced turnaround time for most texts, which allows for rapid input of information into the patient's electronic health record [11]. Data input using SR has been found to be substantially faster than using a keyboard, but is heavily dependent on word recognition accuracy [12]. In addition, the documentation produced by transcription systems is claimed to be more concise, standardized, and maintainable [11]. However, the results were mixed when the total documentation time of speech recognition was compared with other input methods [14]. For the studies covered by Blackley et al. [14], the total documentation time varied drastically following the introduction of SR, ranging from a 92% decrease to a 200% increase. These results highlight that the methods for measuring and comparing documentation times are largely inconsistent between studies, making comparisons difficult [14]. Furthermore, the introduction of new technologies and tools can disturb existing workflows, and potential benefits may not be apparent shortly after adoption.

In general, SR has been found to increase clinician productivity compared to traditional dictation and transcription [14]. In addition, most studies indicate a positive impact when using SR across a variety of devices, such as headsets, mobile devices, and desktops [12]. However, a considerable number of findings are insignificant and the compatibility and successful integration of SR with existing systems warrant further study [12]. Although the general level of user acceptance of SR is high, technical difficulties can significantly affect user experiences [12], which can lead to frustration.

Studies that examined the effect of SR on the quality of medical notes and their use with templates or structured reporting have shown mixed results [14]. Self-edited SR-generated reports were found to contain more errors than those transcribed or edited by professional transcriptionists [14]. In addition, SR can lead to increased documentation costs due to expensive adoption of SR systems or because highly paid clinicians must proofread dictated content instead of transcriptionists [12]. Furthermore, background noise can significantly affect speech recognition accuracy [14], further increasing the need for manual editing.

Assessing the accuracy, documentation times, and costs of SR in clinical documentation is challenging due to inconsistencies in how these factors are evaluated [14]. Training and education are needed to emphasize the need for manual revision of SR-created notes [14]. The role of proper training, support, and maintenance is critical for achieving high accuracy and timeliness with SR [12]. Performance estimates of older systems need to be interpreted carefully, since rapid technological advances may render those results outdated or less generalizable given current AI transcription capabilities [15]. Newer studies describe ambient AI scribes, which summarize and repurpose clinical notes, in addition to transcription [15]. AI scribes are discussed

further in the following section.

2.2 AI scribes

Falcetta et al. [16] describe medical scribes as systems that capture appointment conversations between physicians and patients and are capable of automatically creating documentation based on the interaction. Similarly, Mess et al. [17] describe AI scribes as applications that listen to physician-patient interactions and generate detailed medical notes based on the interaction. Digital scribes record conversations using a microphone and employ an automatic speech recognition (ASR) system that transcribes the conversation into text [16]. Finally, natural language processing (NLP) is used to extract relevant content from transcribed conversation to create clinical notes [16]. This structure is visualized in Figure 1.

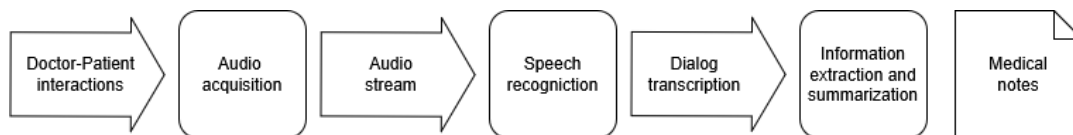


Figure 1: Structure for digital scribes, as described by Falcetta et al. [16]

Natural language processing aims to interpret and manipulate natural language data using different methods and algorithms, depending on the desired outcome [18]. NLP can be divided into two contrasting branches: *Natural Language Understanding* (interpreting language) and *Natural Language Generation* (generating text) [18]. Early NLP approaches relied on rule-based and statistical methods, or machine learning techniques such as Naive Bayes classifiers or Hidden Markov models [18]. However, from 2010 onward the use of various types of neural networks, such as convolutional and recurrent neural networks, enhanced NLP tasks by learning multilevel features [18]. The most recent development in NLP is the emergence of so called "foundational models", which refers to any model that is trained on broad data and is capable of completing a wide variety of tasks [19]. Foundational models are enabled by transfer learning and high levels of scaling [19]. In deep learning, transfer learning is often done by pretraining, where a model gathers knowledge by being trained on a specific task and then applying it to other tasks [19]. While foundational models are not specific to NLP [19], they are still highly relevant due to their application in NLP related tasks. This approach is fundamental to many large language models, which are discussed further in 2.3.

Tierney et al. [20] evaluated the implementation of a smartphone based ambient scribe system. The evaluation found that the use of ambient AI powered scribes decreased the time spent on EHR documentation, especially during after hours (7 p.m. to 7 a.m.)

[20]. This suggests that the system was effective in preventing the accumulation of documentation tasks which would need to be concluded during after-hours periods. In addition physicians and patients reported of improved interaction during visits, due to less time being spent looking at a computer [20]. Furthermore, all patients reported feeling neutral to highly positive about the use of AI during their visit and its impact on the overall quality of care [20]. Overall, the quality of AI created notes was examined to be very high, however there were some cases of hallucination, in situations where summarized content did not fit cleanly to existing note templates [20].

As mentioned, improved patient physician interaction is one of the main benefits pursued by AI scribes. The scribe allows physicians to be more involved and active with patients by removing the computer screen as a focus point, however, physical actions need to be verbalized in order for them to be captured by the system [17]. This is a significant limitation, that can prevent the proposed automatization of using AI scribes and introduce additional tasks to physician workflows, at least in some cases. The objective of increased patient interaction is also highlighted by Falcetta et al. [16]. The authors argue that electronic documentation systems should enable physicians to spend drastically more time interacting with patients than with computers, in order for them to be truly effective [16]. In contrast, Ng et al. [15] note that even if AI transcription systems did not have an immediate effect on patient care, they could indirectly influence clinical outcomes through improved documentation quality and completeness.

However, AI scribes can have some weaknesses. The ability of the AI for distinguishing voices from each other can be limited, and the physician's words could be attributed to the patient [17]. This could result in reduced accuracy, and relevant information missing from created notes. In addition, while some systems can demonstrate high accuracy in specified environments, this does not necessarily mean that their performance is generalizable to a diverse set of scenarios [15]. This can be a challenge, particularly if speech patterns differ, due to accents and different levels of language proficiency [15]. Furthermore, unwanted personal and unrelated conversations may be included as a part of the created note, and need to be manually removed [17], due to the nature of ambient AI scribes. The integration of LLMs may be the next logical step for offering comprehensive AI scribe solutions [15]. This would allow transcripts to be repurposed and summarized, without the need for pristine input accuracy [15].

2.3 Large language models

Generative AI is a branch of artificial intelligence that generates new content based on training data and has great potential to enhance clinical documentation [21]. With the support of natural language processing (NLP), automatic speech recognition (ASR), and prompt engineering, generative AI can transcribe clinical interactions and create notes in structured formats [21]. LLMs are capable of rapidly summarizing and rephrasing information [22]. As a result, repetitive tasks of interpreting and

compressing information that involve little problem solving, such as the creation of discharge summaries, are instructive applications of LLM use in a clinical setting [23]. However, LLMs are not without limitations. LLMs work through probabilistic associations between words, thus they lack the ability to actually understand the user [22]. As a result, they pose the risk of generating outputs that are not based on factual information [5], often referred to as "hallucinations". This raises concerns regarding the use of LLMs in a clinical setting [24]. Regardless of correctness, these outputs are presented with high confidence, potentially misleading the user [5]. Such answers can have serious consequences in a medical setting [5], especially if the user is not aware of these limitations.

The evaluation of the quality of AI generated documentation offers mixed results [24]. LLM outputs are influenced by the complexity and clarity of the input [24] and can vary despite the input remaining the same [17], making it difficult to ensure high quality output. As a result, continuous evaluation of the generated output is required to ensure safe application [24] and medical AI developers should consider ways to make the AI output more uniform [17]. In addition, LLMs can contain biases that are learned by the LLM due to biased training data [5]. For example, the data may under-represent some demographic groups or overemphasize outdated medical practices [5]. Furthermore, LLM training data is often sourced from unverified websites and books [22]. It is often unclear how LLMs generate answers and which parts of the training data influence them [22]. The variation in quality assessment suggests that AI-assisted documentation is not yet consistent across all clinical scenarios and requires further refinement [24].

3 Implications of AI-assisted documentation for privacy and ethics

The use of AI-powered tools for healthcare documentation promises substantial benefits, but also raises important ethical and privacy implications beyond technical performance. These systems process sensitive health data and patient privacy must be protected. Security is essential for protecting data, but in itself, it does not guarantee privacy [25]. Security refers to the confidentiality, integrity, and availability of data, whereas privacy is about the proper and lawful use of an individual's information [25]. Individuals have the right to safeguard their data and decide how and where it is used [25]. Robustness, safety, security, explainability, and fairness are commonly highlighted as key requirements for implementing trustworthy AI systems [7]. These principles combine human-focused ethical values with considerations for legality and regulatory requirements [7]. Issues related to data protection, patient privacy, and the responsible use of AI systems in healthcare documentation are discussed in the following sections.

3.1 Data protection, privacy and robustness

The confidentiality and security of patient information should be ensured through robust regulatory measures, including guidelines for data anonymization, encryption, access control, and protections against misuse by third parties [5]. In addition, AI systems should be robust and resistant to possible attacks, and their use and development must respect the privacy and consent of individuals.

3.1.1 General Data Protection Regulation

The General Data Protection Regulation (GDPR) is an European Union (EU) law that governs the processing of personal data [26]. According to GDPR Article 1, its purpose is to safeguard the fundamental rights of individuals with respect to their personal information while allowing the free flow of data within the EU without unnecessary restrictions. As defined in Article 3, the GDPR applies to all EU based organizations, regardless of where they handle their data. In addition, organizations that handle personal data of people in the EU must follow the GDPR, even if they are not based in the EU [26]. This section provides a high-level overview of the GDPR and its relevance to healthcare documentation. The purpose is not to present the regulation in full legal detail, but to summarize its key principles in a way that supports understanding of their implications for this thesis. Following data protection law is a requirement for all implementations of AI assisted documentation tools, and this introduction to the GDPR creates a foundation for the discussion in the following chapters.

Personal data is defined in GDPR Article 4 as any information that can identify a person, either directly or indirectly [26]. This includes identifiers such as names, identification numbers, location data, or attributes related to an individual's physical, mental, or social identity. In addition, Article 9 defines special categories of personal data, such as data concerning health, for which processing is prohibited unless certain conditions are met. For personal health data, points (a) and (h) in Article 9(2) are particularly relevant. Point (a) permits processing when the data subject has provided explicit consent. Point (h) allows processing when necessary for medical or care-related purposes, such as preventive or occupational medicine, or medical diagnosis. This processing must be performed under the responsibility of a professional bound by confidentiality. [26]

Article 5 of the GDPR establishes six fundamental principles for the processing of personal data [26]. **Lawfulness, fairness and transparency:** Data must be processed in a fair and transparent manner. **Purpose limitation:** Data should only be collected and processed for specified and legitimate purposes. **Data minimisation:** Only data necessary for the stated purpose should be processed. **Accuracy:** Data must be accurate and kept up-to-date. **Storage limitation:** Data should not be stored longer than required. **Integrity and confidentiality:** Data is processed in a manner that ensures the security of the data against unauthorized access, unlawful processing, and destruction. Finally, the data controller shall take **accountability**, by being able to demonstrate compliance with the aforementioned principles [26]. These principles form the foundation of the GDPR, but can also be used as general guidelines when assessing the use of AI in healthcare.

In the GDPR, there are articles that outline how data protection and the security of processing should be ensured [26]. Article 25 introduces the concepts of data protection by design and by default. Data protection by design requires that data protection should be taken into account in the early stages of planning a new way of processing data. Consequently, the data controller must implement measures that enforce the GDPR principles. For example, pseudonymization can support the principle of *data minimisation* by reducing the amount of identifiable information processed. Data protection by default ensures that only the personal data necessary for the specified purpose is collected, stored, and processed. Article 32 sets requirements for the security of processing. Both the controller and the processor must implement measures to protect personal data against destruction, alteration, and unauthorized access. These measures may include pseudonymization or encryption, ensuring the confidentiality, integrity, and availability of processing, being able to restore access to the data in the event of an incident, and having a process for regular testing and evaluation of security practices. [26]

For personal data, processing is allowed only when at least one of the six conditions defined in Article 6(1) of the GDPR is met [26]. These conditions include having consent of the individual, fulfilling contractual obligations with the individual, complying with a legal obligation, protecting the vital interests of the individual,

performing a task carried out in the public interest or under official authority, or pursuing the legitimate interests of the controller, provided these do not override the fundamental rights and freedoms of the individual. Article 7 of the GDPR defines the requirements for individual consent. Firstly, when processing is based on the consent of an individual, the data controller must be able to demonstrate that the individual has consented to the processing. Secondly, the request for consent from the individual must be presented in plain language and in an unambiguous way, so that the individual understands what they are consenting to. In addition, the individual has the ability to withdraw their consent at any time, and withdrawing consent should be as easy as giving it. Finally, consent must be freely given and cannot be conditional on processing personal data that is not necessary for the performance of a contract or service. [26]

Personal data can be processed by multiple companies or organizations depending on the use case. To address this, the GDPR defines two key roles involved in processing: the data controller [Art. 4(7)] and the data processor [Art. 4(8)] [26]. The controller determines the purposes and means of processing, whereas the processor acts on behalf of the controller and processes data according to its instructions [26]. For healthcare, an example of a data controller could be a nursing home, since it determines how patient data can be used and for what purpose. A company that provides a digital EHR system to the nursing home would generally be considered a data processor, as it processes personal data on behalf of the nursing home and according to its instructions.

According to Article 24, the data controller is responsible for implementing and demonstrating that appropriate measures are in place in order to comply with GDPR requirements [26]. In addition, the controller must maintain records of processing activities, including the purpose of processing, categories of data subjects, and a general description of technical and organizational security measures. Furthermore, according to article 35(1), the controller must conduct a Data Protection Impact Assessment (DPIA) in cases where processing is likely to result in a high risk to the rights of individuals, for example, when using new technologies [26]. Based on this requirement, it is reasonable to conclude that the development of AI-assisted documentation systems in healthcare will, in most cases, constitute high-risk processing, as the AI system will handle sensitive personal data. Therefore, conducting a DPIA is necessary in such cases.

As mentioned above, the processor processes data on behalf of the controller. As defined in GDPR Article 28, data can be processed only in the manner that the controller has defined, and the processor must be able to guarantee appropriate technical and organizational measures to comply with the GDPR [26]. In addition, the processing is governed by a binding contract between the processor and the controller. The contract specifies the nature and duration of the processing, its purpose, and the types of personal data that are processed. Furthermore, the processor cannot involve another processor without the consent of the controller. If the processor engages another processor for specific data processing activities, the same conditions outlined

in the original contract with the controller must also apply to the sub-processor, and these obligations must be defined in a separate contract between the processor and the sub-processor [26]. For example, if an EHR supplier intends to integrate a third-party AI service into its application, it must first obtain the consent of the data controller and establish a contract that clearly defines the processing terms.

In addition to regulations regarding the controller-processor relationship, rights of the individual are defined in articles 15 to 22 [26]. These rights should be taken into consideration when AI powered documentation systems are used and developed. Many of the rights have conditions and exceptions that will not be listed here, but the rights can be summarized as follows:

- **Right of access by the data subject** (Art. 15): Individuals have the right to know if their personal data is being processed and to obtain a copy of the data. In addition, they have the right to know the purposes, methods, and categories of personal data processed.
- **Right to rectification** (Art. 16): Individuals have the right to have inaccurate or incomplete data corrected.
- **Right to erasure** (Art. 17): Individuals have the right to request the deletion of their personal data.
- **Right to restriction of processing** (Art. 18): Individuals have the right to limit the processing of their personal data, even if the data is not deleted.
- **Notification obligation regarding rectification or erasure of personal data or restriction of processing** (Art. 19): If data is corrected, erased, or its processing is limited, the controller must inform the individual and any recipients of the changes.
- **Right to data portability** (Art. 20): Individuals can request their personal data in a structured, machine-readable format and have the right to transfer it to another controller.
- **Right to object** (Art. 21): Individuals can object to the processing of their data for certain purposes, as long as the controller has no overriding reason to process the data.
- **Automated individual decision-making, including profiling** (Art. 22): Individuals have the right not to be subject to decisions made by automated processing.

The GDPR serves as a key regulatory framework for protecting individuals and their personal data. For AI-assisted documentation systems in the healthcare sector, compliance with GDPR rules and principles is essential to ensure that sensitive health

data is processed lawfully and ethically. Failure to comply can lead to severe penalties, including fines of up to 20 million EUR or 4% of a company's global annual turnover [Art. 83] [26]. Additionally, supervisory authorities have the power to impose corrective measures, such as ordering the suspension of data processing [Art. 58] [26]. The requirements for explicit and informed consent, along with the individual's rights to track and request removal of their data, shift the power balance towards the patient in determining how their data is used [27]. These rights also highlight the importance of protecting patient privacy and ensuring appropriate data ownership governance [27].

Although GDPR applies to all EU-based organizations and those processing the data of individuals in the EU, countries outside the EU have their own data protection laws. For example, Switzerland enforces the Federal Act on Data Protection (FADP) [28]. A detailed analysis of the FADP is not necessary for the purposes of this thesis, and it is sufficient to note that the regulations and principles of the act are similar to those of the GDPR.

3.1.2 Security, robustness and privacy preserving techniques

As healthcare data is highly sensitive in nature, strong safeguards should be in place when implementing AI-powered documentation systems. The implementation of AI-powered services in healthcare is often managed by external commercial entities [4]. Machine learning (ML) models can be deployed as cloud-based services [29], which means that data must be transferred to third party servers. As a result, these companies play an increasingly significant role in accessing, processing, and safeguarding patient health information [4]. Therefore, adequate safeguards are needed to maintain privacy and patient agency [4]. There exists a tension between ethically providing medical care and generating a profit, which AI technology providers are not immune to [27]. Corporations may be tempted to monetize or otherwise gain from personal data, especially if such behavior is not disincentivized by severe enough legal penalties [4]. In addition to data protection, these systems should be reliable and resilient against irregularities and attacks. Ensuring the robustness of ML models remains a significant challenge, particularly in security-critical applications [3]. The literature suggests that ML systems are often not safe, secure, and robust, as they can be manipulated through various types of attacks [7]. Despite different proposed defense strategies, ML models remain vulnerable to data corruption, distributional shifts, and malicious threats [3].

Attacks against healthcare data sources or related AI models are a significant privacy concern, because they can allow unlawful access to sensitive information, such as medical records or personally identifiable data [3]. Khalid et al. [3] divide attacks against ML-based systems into two categories: Attacks against data (data privacy) and attacks on the model (model privacy). Attacks against data include re-identification, reconstruction and property inference [3]. Re-identification attacks lead to the identification of individuals through data that has been de-identified [3].

Reconstruction attacks result in the reconstruction of the original raw data, including private data, through the use of non-private recognizing information [3]. Property inference attacks allow the attacker to identify dataset characteristics that were not connected or intended by the task at hand, such as the ratio of women and men in a healthcare dataset [3]. When real-world healthcare data is used to train AI models, these types of attacks can result in privacy violations, even if the data has been anonymized. Attacks against AI models include membership inference, model inversion, and adversarial machine learning attacks [3]. The goal of a membership inference attack is to deduce whether a particular input was part of the model's training data by examining model outputs [3]. Similarly, in a model inversion attack, the attacker aims to reconstruct parts of the training data [3]. In adversarial machine learning attacks, the attacker seeks to reduce the accuracy of the model or manipulate its output by introducing subtly misleading data during training [3]. Similarly to attacks against data, membership inference and model inversion can reveal private information, but direct access to the training data is not required. Furthermore, the performance of the model can be compromised or manipulated by corrupting the training data, potentially leading to inaccurate documentation and errors in care.

In the literature, many different defense strategies have been proposed against particular types of attacks targeting cloud-hosted machine learning models [29]. For example, Salem et al. [30] present dropout and model stacking as potential defense strategies against membership inference attacks. The dropout method randomly deletes a proportion of edges during each training iteration in a fully connected neural network, and model stacking organizes multiple ML models in a hierarchical way [30]. These methods are typically used to prevent overfitting, but have been shown to greatly reduce the effectiveness of membership inference attacks, without compromising model accuracy [30]. However, many proposed attacks and defense strategies do not take into account adaptable adversaries [29]. In practice, attackers continually improve their attacks based on knowledge of current defense systems, and defenders aim to develop defenses that are effective long-term [29]. Furthermore, the proposed defenses are limited to particular settings and attacks, and therefore are not generalizable [29]. As a result, the development of adversarially robust ML models remains an open field of research [29].

Since many defense strategies remain context-specific, researchers have explored privacy-preserving machine learning. For ML models trained on user data, it is desirable that they do not learn information that could violate individual privacy [29]. The dependency of AI algorithms on high-quality learning data raises further concerns in healthcare, where sensitive information is at risk [3]. Data leakage or misuse can cause significant harm to patients, healthcare providers, and software vendors [3]. Therefore, privacy-preserving machine learning techniques have been proposed to enhance privacy and strengthen user trust [3].

Khalid et al. [3] present several privacy-enhancing techniques, some of which could be relevant to the implementation of privacy compliant AI-assisted documentation tools

and related AI models. Homomorphic encryption is a technique in which the data owner encrypts the data before it is sent for computation [3]. As a result, collaborating with third parties and performing cloud computations is safe, since the data is not revealed to outside parties [3]. However, in practice, complicated homomorphic encrypted algorithms create significant computational overhead and are slow to use [3]. Differential Privacy (DP) aims to facilitate data anonymity by introducing noise into the data [3]. DP is best used for queries where slight alterations of data do not have a significant effect on the result [3]. However, if the data is highly sensitive to change, the need to add large amounts of noise limits the working of the algorithm and may result in decreased accuracy [3]. Federated learning (FL) is a technique that aims to create a ML model with data from different sources, without the need to transfer data away from the source [3]. The model is trained locally on each source, and the difference is compared to a central model, which is then updated [3]. However, FL can be prone to attacks and the cost of communications between clients and sender is a primary obstacle to FL scalability [3]. There are many different proposed techniques for preserving privacy in ML models, but most of them are still vulnerable to different kinds of attacks or are not feasible in practice [3]. Nevertheless, the aforementioned privacy preserving techniques have potential in creating ML models for healthcare documentation while preserving privacy. It should be noted that the utilization of privacy preserving techniques often has a direct negative effect on model performance [3]. As a result, the balance between privacy and accuracy must be carefully examined to protect the privacy of users and ensure the proper function of the application [3].

3.1.3 Consent

Acquiring informed consent from a patient is a practical application of respecting people and their autonomy [7]. Furthermore, patients have expressed that they wish to have the option of opting out of the use of AI tools in their care [31]. Using AI tools in the care of patients who have explicitly prohibited the use of AI could lead to a serious breach of trust [31]. In addition, some patients may feel misled or betrayed if AI is used without disclosure, as some of the work typically done by humans would be performed by computers instead.

As mentioned, article 7 of the GDPR defines guidelines for obtaining the consent of an individual with respect to data processing [26]. In particular, the request for consent is defined as being presented in clear and unambiguous language, so that the individual understands what they are consenting to [26]. However, the black-box nature of AI systems makes it challenging to acquire informed consent, as it may be difficult for patients and clinicians to fully understand how the AI system functions [7]. Therefore, improving the explainability of black-box models facilitates the acquisition of informed consent [7]. At least, patients should be informed about the black-box nature of the AI system and be aware of the possible benefits and drawbacks [7].

The acquisition of consent requires collaboration from healthcare professionals and

patients. Tierney et al. [20] described how physicians were educated on how to obtain consent from patients for using an ambient scribe tool. As a result, a standardized EHR template was developed to document consent acquisition. In addition, patients received educational handouts about the use of the tool and it was clearly defined in which areas the tool could be used. [20]

Although obtaining consent for data processing is both a legal requirement and an ethical best practice, alternative approaches to data sharing have also been proposed. Panagopoulos et al. [32] propose a more lenient approach to the use of data by AI companies. Their proposed model of collective data management (CDM) is similar to the collective rights management (CRM) used in the copyright sector. In CRM, a central organization acts on behalf of content creators to license, monitor, and collect royalties for the use of copyright-protected material, such as music or literature. Similarly, CDM would allow the data of participating individuals to be used more freely, while a management organization would be responsible for monitoring its use and negotiating compensation accordingly. The authors acknowledge that the implementation of CDM would require considerable changes in legal contexts and public attitudes, in addition to further practical experimentation. Nevertheless, the model seeks to address the trade-off between data protection and data availability by encouraging broader data sharing to fuel AI innovation, while still offering individuals a share of the value generated from their data contributions. [32]

3.2 Responsible AI use in healthcare documentation

Transparency, accountability, and fairness are areas that need to be addressed for ethical use of AI in healthcare [5]. In the context of this thesis, the responsible application of AI refers to building trust in the technology among users and subjects, along with ensuring adequate awareness of its capabilities and limitations. In addition, it also involves accountable use, where patients are protected from potential errors and results are ensured to be fair and free of bias. The following sections examine how these characteristics can be supported in the context of AI-assisted healthcare documentation.

3.2.1 Transparency, explainability, and trust

Recent literature highlights the importance of using healthcare AI systems in a transparent manner. Users and subjects should also have a high-level understanding of how the AI tool functions and what role it plays in the documentation process to facilitate trust and responsible use. The full potential of AI in healthcare is hindered by a lack of transparency, which prevents clinicians from fully trusting these systems [7].

Richardson et al. [31] examined patient attitudes toward using AI in healthcare.

In general, it was found that participants were enthusiastic about the use of AI in healthcare, as they felt it would offer new opportunities to heal as many people as possible. However, even though the participants were not deeply informed on the technical intricacies of AI, they were able to identify concerns that could negatively affect them or their families. It was stated that cautious implementation, careful transition periods, and regulatory oversight are needed to protect patients, as AI is an emerging technology. In addition, increased healthcare costs have been raised as a concern, due to perceived high development and deployment costs of such systems. [31]

Bracken et al. [24] also found that the experiences and attitudes of healthcare professionals toward AI-driven documentation were generally positive. Users appreciated the efficiency and structured formatting achieved by these AI tools, but remained skeptical about their reliability and the possibility of losing narrative detail in AI-generated clinical notes. [24]

Due to complexity of AI technologies, many AI systems are "black boxes" by nature. As a result, healthcare professionals and patients are not aware of how AI works or why it produces certain results. These black box systems reduce trust in AI systems [6]. The development of transparent AI systems that are understandable and explainable to a wide range of users, such as healthcare professionals and patients, aids in building trust, informed decision-making, and responsible and ethical use of AI systems [6]. If the reasoning of a system can be explained, humans can evaluate whether the reasoning is valid [27] and decide whether to trust it or not. If the reasoning cannot be explained, this opportunity does not exist [27].

Rasheed et al. [7] highlight the importance of explainable machine learning (XML) in healthcare. XML helps visualize the relationship between potentially biased features and the outcomes of models, thus providing a fair opportunity to analyze model architecture and learned parameters. In addition, XML can help researchers and developers in picking the best model. Furthermore, the explainability of ML models allows researchers and end-users to participate in improving the models and better trust their outputs. [7]

In addition to aiding with ethical challenges, having the ability to explain the results helps end users (healthcare professionals, patients) who want to learn, understand and manage ML algorithms efficiently [7]. However, the explanation methods are not standardized, and it may be difficult for untrained clinicians to interpret the results [7]. Therefore, there should be a platform for connecting medical professionals with XML researchers, in order to create a standardized way of explaining the results [33]. Another challenge is being able to explain the results to non-technical people, such as patients, to increase their trust in the system [7].

Although XML holds promise, highly technical explanations may not be practical in a healthcare setting. Nevertheless, the existence of XML is a positive for the transparency of AI and can be taken advantage of if people with appropriate technical

proficiency are present and able to "translate" the explanations to people who are not familiar with XML techniques. A more hands-on approach to increasing AI transparency could be better suited for healthcare documentation settings. Mess et al. [17] advocate for medical AI developers to develop ways to draw the attention of physicians to areas where AI has low confidence, for example, by highlighting text or showing a confidence score in percentages. As a result, these areas could be manually inspected and corrected if needed.

Trust building, proactive innovation, and recognition of complex social factors are needed to address patient apprehensions about AI in healthcare [31]. Deliberate patient engagement is also essential to prevent further skepticism from emerging [31]. In addition, healthcare professionals should have at least a high-level understanding of the AI-powered documentation systems they use to ensure safe and effective application.

3.2.2 Accountability and the role of AI in healthcare documentation

Patients interviewed by Richardson et al. [31] emphasized that physician oversight of AI systems is critically important. The participants felt that clinicians should remain responsible for patient care and protect patients from potential mistakes made by AI [31]. It was also expressed that AI should not have the ability to act autonomously in care monitoring or clinical decision-making [31]. Although AI-powered documentation tools may not directly influence treatment decisions, they can still have an indirect impact, for example, by introducing errors into clinical notes that affect care planning. Therefore, the expectation that healthcare professionals remain accountable for the actions and errors of AI systems should be considered when evaluating responsible use of AI-assisted documentation in healthcare.

In addition to the expectation of physician supervision, participants expressed concerns about healthcare professionals and the technologies they use becoming too dependent on AI [31]. They also raised concerns about a decline in care quality in situations where AI tools are suddenly unavailable and healthcare providers are overly reliant on them [31]. This highlights the need to treat AI as an assistive tool, rather than a replacement for existing practices. To support responsible use, healthcare professionals should have a clear understanding of how AI-powered documentation systems influence patient care. If users assume that the AI will perform flawlessly or handle tasks independently, it becomes more difficult to assign accountability when errors occur.

Many studies related to human-AI collaboration focus on the design and implementation of AI systems, rather than the evaluation of systems already in use [34]. Bossen and Pine [34] examined the use of a useful but error-prone AI tool used by clinical documentation integrity specialists, which uses natural language processing to automatically suggest codes based on patient documentation. This real-world case offers several lessons on the conditions under which AI tools can be successfully adopted in clinical environments. Specifically, the study highlighted the following five factors that

supported its successful design, implementation and use [34]:

- Flexible integration into existing workflows
- Augmentation of human work, rather than replacement
- Retention of human control and decision-making
- Transparency and explainability of system actions
- Support for user-driven experimentation and learning

It would be reasonable to hypothesize that the primary factor determining the success of an AI-based documentation system in a healthcare setting could be its accuracy and ability to avoid errors. However, based on the five factors identified in [34], this is clearly not the case, at least in this particular context. The AI system is utilized as a supportive tool, employed when convenient for the user. It does not disrupt existing workflows and is not designed to automate tasks entirely, but rather to assist with them. Users are aware that the system may produce errors, but human oversight and control are maintained throughout its use. This combination of factors contributes to a successful human-AI collaboration, even though the tool is prone to errors.

However, as stated by Bossen and Pine [34], there are some limitations to these factors. The collaboration between the AI system and its users had continued for a long time, which allowed users to become accustomed to the system and the mistakes it makes [34]. Such long adjustment periods may not be possible in high-stakes environments, where the results of the system directly impact patient care [34]. As a result, the use of inherently error-prone systems is limited by the potential consequences of their application and the need to validate their outputs. In clinical documentation, this means that such systems may be acceptable for low-risk tasks but unsuitable for use cases where even minor errors can cause serious harm. Nevertheless, users of the AI system were able to remain fully accountable for its performance, as they understood how it functions and used it as a tool to support their work.

All stakeholders should be actively involved in the AI implementation process in medicine [27]. Especially healthcare professionals should be supported when new AI-based documentation tools are introduced. Key obstacles to adoption include lack of familiarity with new technology, difficulties in access or activation, and poor integration with existing workflows [20]. Tierney et al. [20] describe how physicians were supported and trained to take advantage of a new ambient scribe tool. Physicians were introduced to the tool through an interactive question-and-answer webinar, and training sessions were held to ensure the effective and safe use of the tool. In addition, internal support sites and staff helped install and use the tool when required. Beyond technical support, healthcare professionals should also be educated on the potential benefits and limitations of AI [27]. Ideally, they would have technical knowledge about the underlying algorithms and datasets and their effects on outputs, but this may not be

a reasonable expectation due to finite resources and busy schedules [27]. Nevertheless, the key to implementing AI-driven medical scribe solutions lies in supporting the use of new tools, fostering a culture of innovation, and gradually integrating these tools into existing workflows [20]. In addition, it was noted that continued monitoring and proactive assessment of new tools is necessary [20].

3.2.3 Fairness

The healthcare domain is filled with medical terminology, abbreviations, and context-dependent language [21]. Therefore, in order for generative AI systems to achieve high performance, they need to be trained on diverse and comprehensive data, containing a wide range of medical conditions, patient demographics, and clinical scenarios [21]. AI regulations could require the use of unbiased and diverse training data to ensure that existing disparities in the medical field are not maintained or amplified in the output of the model, thus promoting fairness in the delivery of healthcare services [5]. The quality of training data is cited as a concern by patients [31]. According to Bell et al. [35], roughly 1 in 5 patients that had read 1 or more of their medical notes in the last 12 months reported a perceived mistake, of which 42.3% were considered serious. The data of the real-world EHR system can be inaccurate or even false in some scenarios, and the use of such data for AI development raises concerns about model performance [31].

Discourse regarding the use of AI in healthcare should include a diverse set of stakeholders, such as technologists, healthcare providers, ethicists, and patients [6]. Engagement of different perspectives ensures that AI development is based on real care needs, promoting patient-centered innovation [6]. This approach ensures that AI serves the needs of the healthcare community and respects the values and dignity of individuals [6]. Biswas and Talukdar [21] highlighted that the balance between technological innovation, patient-centered care, ethical principles, and regulatory compliance is key to successfully harnessing generative AI in healthcare.

4 Data Ethics Decision Aid (DEDA) framework

Data Ethics Decision Aid is a framework for increasing awareness, generating discussion, and documenting the decision making process regarding data and AI projects [36]. DEDA was developed as an iterative process through 2016 to 2018 by Utrecht Data School and Utrecht University. The DEDA framework was originally designed to assess the societal impact, values and responsibilities stemming from data based public management. As a result, DEDA has been used by various public-sector organizations in the Netherlands, such as municipalities and the Ministry of General Affairs. [36]

DEDA was developed using an empirically driven research method that focused on openness about collected data through interviews, focus groups, and surveys [36]. The authors cite ethnographic approaches to follow actors [37, 38] and participatory action research (PAR) [39, 40] as a partial inspiration for the method. The purpose of this approach was to recognize how the framework could be integrated into existing workflows [36]. As a result, two main goals were identified: increasing awareness of ethical pitfalls in data projects and making the related underlying values more visible [36].

As mentioned, the DEDA framework was developed as an iterative process [36]. The first prototype was developed based on expert interviews, in which people with expertise in the field of data-driven public management were interviewed and common ethical issues were identified. After that, feedback on the prototype and the following versions was gathered through three iterations of focus groups, which contained researchers and employees of municipalities. Finally, the framework was refined based on feedback that was gathered during the use of the framework in real data projects. Notably, ethical pondering of potential issues was deemed as theoretical and difficult for participants with no relevant education to understand, thus a more practical approach was defined. [36]

The authors of the DEDA framework reference three areas of data ethics by Floridi and Taddeo [41] (ethics of data, algorithms, and practices) as considerations for understanding how practices, algorithms, and data life-cycles function in the municipality context [36]. In addition, value-sensitive design and ethical pluralism formed the basis for the research on the framework [36]. The use of value-sensitive design allowed the authors to account for the complexity of ethical research, place a focus on the explicit mapping of inherent values of technology, and highlight the importance of connecting designers of systems with other stakeholders [36]. The use of ethical pluralism meant that instead of considering a particular area of ethics as definitive, a more open approach accounts for the effects of personal context and different types of ethical reasoning [36].

The framework has several motivations and objectives. First of all, it can be difficult to explicitly define ethical challenges of technical implementations, even if it may be

easy to see that they exist [36]. In addition, these challenges cannot often be solved by applying relevant laws and regulations, since some data uses might be legal, even though they are immoral. Furthermore, many already existing tools for assessing these issues are too abstract in nature and do not take into account the context of the project they are assessing. The authors claim that the use of a framework like DEDA will facilitate discussion and debate about data ethics. In addition they claim that ethical guidelines help in explaining possible ethical risks to the public and aid those working with data to enhance their understanding. Furthermore, laws and regulation naturally lag behind innovation, and ethical guidelines can be used in such cases, where formal rules are not yet in place. [36]

The effectiveness of the framework was evaluated through a survey that participants completed before and after using the framework [36]. The survey questions were based on five categories of knowledge as defined by Bloom et al. [42]: remembering, understanding, applying, analysing, and evaluating. The survey results showed, with an acceptable level of reliability, a self perceived increase in all five categories, most notably in "remembering" and "evaluating" [36]. These results and the widespread use of DEDA in the Netherlands makes it an appealing tool for assessing ethically complex data projects. In addition, preliminary results have indicated that DEDA is a useful tool outside the Dutch governmental context [36].

II. Case Study: Smart Care prototype evaluation

5 Methods

This thesis used a qualitative case study as the research method. The case study consisted of two main components:

1. The definition, design and implementation of the Smart Care prototype.
2. Evaluation of its ethical and privacy-related implications using the Data Ethics Decision Aid (DEDA) [36] framework.

According to Yin [43], the case study method can be used to gain an in-depth understanding of a real-life phenomenon in its specific context. The ethical evaluation of artificial intelligence use in a healthcare setting is a highly complex topic that involves many interacting stakeholders. Since the study is approached from the perspective of evaluating the implications of a new technical implementation, reliable quantitative data was not available. Therefore, a qualitative approach was chosen that focused on subjective experiences, stakeholder perspectives, and ethical considerations that cannot be easily measured or quantified.

The DEDA framework has two main goals for its usage: increasing awareness and understanding of the complexity of data ethics, and encouraging structured dialogue around ethical aspects in data projects [36]. The framework is not intended as a checklist or formal impact assessment, but rather as a tool for open-ended discussion on topics such as algorithms, data sources, privacy, and responsibility [36]. It was chosen for this thesis as a guiding resource for the ethical evaluation of the Smart Care prototype, as it allows stakeholders to discuss, reflect on, and evaluate the project from different perspectives early in its development, helping to identify potential concerns in advance.

The DEDA evaluation was conducted as a semi-structured group interview, with participants from three relevant stakeholder organizations: Vitec Raisoft, BESA QSys, and Adamcares. Each company was represented by a participant who has been involved with the Smart Care project and was able to discuss the questions of the DEDA assessment. The evaluation mainly followed the structure of DEDA REMOTE V2 [8], consisting of a project overview, a discussion of stakeholder values, data-related and general considerations, and a conclusion. The evaluation was organized as an online meeting and the author of this thesis acted as a moderator. The moderator facilitated the discussion, provided clarifications regarding the assessment process, and documented the results. During the assessment, open discussion was encouraged and each participant was asked to reflect on the questions from their own perspective to capture a comprehensive view of the project and its implications. The discussion was recorded so that the responses could be examined in full after the interview. Due

to time constraints, the moderator answered the concluding questions (presented in [subsection 7.4](#)) after the workshop, based on the workshop discussion. Subsequently, these answers were reviewed and approved by all participants.

6 The Smart Care prototype

This section provides an overview of the Smart Care prototype, which includes project goals, system requirements, collaborative partners, AI integration, and implementation results. The purpose of this section is to provide the reader with sufficient context to understand the evaluation presented in [section 7](#).

6.1 Project background and stakeholders

The Smart Care prototype was developed as part of this thesis to explore the potential of AI-assisted speech recognition in healthcare documentation. The prototype integrates speech-to-text, AI-based text structuring, and context recognition functionality into Vitec Raisoft's existing documentation system, ePDoc.net. The prototype was developed in collaboration with Swiss based companies, BESA QSys AG and Adamcares AG.

Vitec Raisoft Oy is a Finnish software company that provides digital solutions focused on healthcare, rehabilitation, and mental health [44]. Founded in 2000 and acquired by the Swedish Vitec Software Group in 2022, the company develops tools that support care planning, quality monitoring, and resource management. Many of these solutions are based on the internationally recognized interRAI assessment system, developed by the independent research network interRAI. One of Raisoft's core products is RAIssoft.net, a web-based Software-as-a-Service (SaaS) platform that facilitates structured data collection and reporting. The system gives clinicians and managers access to real-time information, allowing for effective decision making. RAIssoft.net can be integrated with existing EHR systems and adheres to international quality and data security standards. [44]

In addition to RAIssoft.net, the company also offers ePDoc.net, a digital care documentation tool. ePDoc.net is designed to support nurses in documenting care needs, interventions, and care quality in a structured and transparent way. The system organizes documentation into modules, such as the care report, measurements, care plan, and medications. It is also fully integrated with RAIssoft.net and can be accessed through the same web-based interface. [45]

BESA QSys specializes in IT-based solutions for assessing care and nursing needs in the health and social care sector [46]. Services are aimed at institutions in long-term care, social services, and disability services, as well as day care centers seeking a professional care needs assessment. The company emphasizes integrated care and ensures that its solutions are aligned with current nursing science, legal requirements, and the evolving needs of healthcare providers. [46]

Adamcares develops AI-powered solutions for automating healthcare documentation [47]. The system uses Natural Language Processing to convert spoken communication

into text, followed by Retrieval Augmented Generation to optimize the transcription by correcting terms, abbreviations, dialects, and accents [47]. It is designed for easy integration via an AI API, which returns content in a format compatible with patient records [47]. The AI service used in the Smart Care project is provided by Adamcares.

6.2 Goals

The short-term goal of the Smart Care prototype was to develop a proof-of-concept application to evaluate the integration of speech recognition and AI-based text structuring within the existing ePDoc.net system. The prototype is intended for internal testing and future presentation to selected ePDoc.net customers in order to gather preliminary feedback. Based on the results of testing and stakeholder input, a decision will be made about whether to proceed with further development.

In addition to the short-term goals, several long-term goals were identified. Although these are outside the scope of the prototype implementation, they are provided here to provide context and motivation for the project. For nurses, the goals include reducing the time spent on documentation, thereby freeing up resources for better caregiving. Furthermore, the use of AI-based text structuring aims to improve documentation quality. Moreover, adding support for multiple languages would allow nurses with limited language skills to contribute to documentation more easily.

6.3 System requirements

For the Smart Care prototype, no formal requirements elicitation process was conducted and the requirements were not explicitly listed prior to development. However, an understanding of functional and non-functional requirements was derived from BESA QSys-provided descriptions of the use case and UI mock-ups, as well as discussions within the development team about the solution. The following requirements guided the development process.

Functional Requirements

1. The user can log into the Smart Care application using their ePDoc.net credentials.
2. The user can select a resident for whom the documentation is to be created.
3. The user can start and stop recording their voice.
4. The application can transmit recorded voice input to the AI service.
5. The AI service processes the audio input as follows:

- (a) Supports input in (Swiss) German.
 - (b) Identifies the documentation module to which the content belongs.
 - (c) Returns output in a predefined structured format (JSON).
 - (d) Restructures free-text content to improve clarity, style, and readability.
6. The application pre-fills a documentation form according to the AI response.
 7. The user can accept, modify, or discard the form.
 8. The user can save the form contents to ePDoc.net.

Non-Functional Requirements

1. The application must be mobile-friendly.
2. The application must support common mobile platforms (such as Android, iOS) and web browsers (such as Google Chrome, Safari).
3. The application should be easy to use and require minimal user interaction.
4. The application should provide appropriate feedback to the user during long API call(s) (up to 30 seconds).

The scope of the prototype implementation was restricted to one ePDoc.net module called Care report. Therefore, the application contains only one input form, and the AI only considers one module as an option. However, the system requirements were still fully considered in a way that development can be easily continued beyond the prototype and extended to additional modules.

6.4 AI service integration

The AI service used in the Smart Care prototype was provided by Adamcares. Data processing, including transcription and response structuring, is performed on remote servers, and the cloud infrastructure provider is ISO/IEC 27001:2022 certified. The service is accessed through an API that accepts a base64-encoded voice recording as input and returns a structured response based on the transcription and available modules. For the Smart Care prototype, only the Care report module was implemented on the AI service side.

For the AI service, a response structure was defined to specify the format in which relevant information gathered by the AI from the voice recording would be returned to the application. The structure was defined in JSON format and closely mirrored the fields of the care report submission form in ePDoc.net. For example, the main dictated

content of the report was returned as a string, and a reference to a specific interRAI instrument question could be included if the content contained a specific code from a predefined list. Fields in the structured response were populated if recognized by the AI from the transcribed voice recording. Initially, most of the input fields were included in the response. However, some fields required selecting values from lists that are dynamically defined by each organization using ePDoc.net. Technically, it would have been possible to configure the AI service with organization-specific lists, allowing it to choose values based on the requesting organization. However, the effort and resources required to add and maintain these lists outweighed the potential benefits. Consequently, support for dynamic lists was removed from the AI response. These fields must now be manually selected by the user or left at their default values. Upon further evaluation, it was concluded that pre-filling these fields was not essential for the most common use cases of the Care report, and manual selection was not considered a serious limitation.

6.5 Implementation results

The Smart Care prototype was implemented within the existing ePDoc.net codebase, which offered several advantages. Firstly, users can log into Smart Care using their existing ePDoc.net credentials, as the same authentication system is used. Secondly, logging and access control systems are already in place and reused in this application. User events, such as accessing pages, fetching residents from the database, calling the AI-API, and submitting forms, are recorded in the system. Furthermore, access to Smart Care is limited to users who have been explicitly granted the required permissions, and access to the Care report module is also necessary. In the future, if additional modules are supported, the system and AI will only consider those modules to which the user has access.

Smart Care was developed as a single page application, primarily intended for use on mobile devices. Its features closely follow the requirements described in 6.3. The user selects a resident, dictates documentation using voice input, and reviews a form pre-filled by the AI response. The data flow between the user, Smart Care, the AI-API, and ePDoc.net is illustrated in 2, and it has the following steps:

1. The user records their voice in the Smart Care application.
2. The recorded voice is sent to the AI-API.
3. The AI-API returns a structured response.
4. A documentation form is pre-filled and displayed to the user.
5. The user accepts, modifies, or discards the form.
6. Accepted or modified form contents are saved to ePDoc.net.

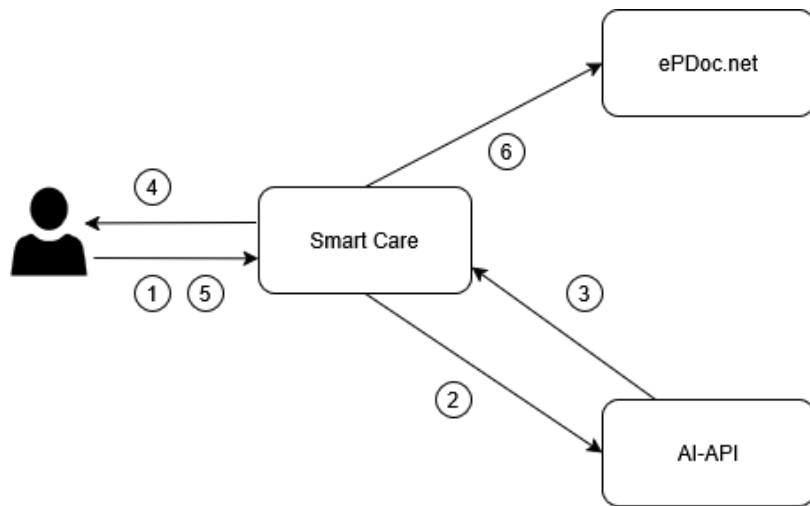


Figure 2: Smart Care application data flow

7 DEDA assessment results

The following sections present the results of the DEDA evaluation, based on the discussion held at the stakeholder workshop. The results are presented in the same order as they appear in the DEDA assessment. The full list of questions is available in appendix A. The results are presented as is, without additional input from the thesis author. The results are discussed and compared to the overlying literature in section 8. When necessary, brief clarifications are included to explain how certain topics were approached in the context of the Smart Care project. In some cases, the discussion extended beyond the scope of the Smart Care prototype. These results are still included, as they help form a more comprehensive understanding of the project and highlight potential future considerations.

7.1 Project overview and values

In the first phase of the assessment, the Vitec Raisoft representative presented a brief overview of the Smart Care project. This overview was based on nine questions that were answered before the workshop, covering project goals, data usage, stakeholders, pursued benefits, and potential concerns.

The primary goal of the project was to create a proof-of-concept Smart Care AI application. During the development of the application, the objective was to gather insights into the AI-API, system requirements, and design specifications. In addition, the prototype was intended for deployment and evaluation with selected test customers. Three main types of data used in the prototype were identified: user speech input, approved or modified form data submitted to ePDoc.net, and data used to train the AI models used by the AI provider. Three stakeholder groups were identified, each with pursued benefits: nursing staff (reduce time spent on documentation), managers (improve documentation quality), and organizations (free up resources for better caregiving). For the participating companies, the pursued benefits included increased revenue from upselling to existing customers and attracting new customers and markets. Potential concerns with the project included legal and data security issues that could arise from integrating a third-party AI solution that processes patient data.

In the second phase of the assessment, each participant listed important values for their organization. Values such as trust, transparency, simplicity, fair pricing, and purposeful innovation were highlighted. These values would then be reflected on when discussing the questions.

7.2 Data-related considerations

Algorithms

The Smart Care prototype does not necessarily contain algorithms in the traditional sense. However, for the purposes of the DEDA assessment, the AI system was considered as an algorithm to facilitate discussion. In addition, false negatives and false positives were interpreted as instances where the API produces unexpected or incorrect output.

The discussion addressed how the AI system functions and whether it is important for someone from the team to be able to explain what it does. It was considered beneficial to be able to describe the system on a high level, particularly in situations where a customer wishes to know how the system works. The representatives of Vitec Raisoft and BESA QSys demonstrated a general understanding of the system.

The representative from Adamcares gave a detailed explanation of the system, which is built on an agent-based framework that selects the most suitable models for different use cases. The process begins with transcribing the user's voice input into text. This transcription is then refined by various agents, for example, by replacing words or restructuring sentences. Manager agents analyze the content to determine to what predefined module(s) the input is related to, and the task is divided among agents accordingly. Each module has a predefined output format to ensure compatibility with the integrated system. In general, each agent performs a specific task and leverages appropriate neural networks (LLMs or SMLs), as well as prompt engineering and Retrieval Augmented Generation (RAG).

It was noted that the correctness of the AI output is decided by the user. The user has full control to accept, modify, or discard output, through the pre-filled form. In conjunction, it was deemed that the project is not aiming to replace existing documentation practices entirely with speech input, but rather to make existing practices more efficient. However, the possibility of speech-only documentation was seen as a future consideration, as long as the trust in the AI system is high.

A recurring topic was the possibility of improving the AI service in the future. One proposed approach involved comparing transcriptions of user speech input and final form data saved to ePDoc.net. The output of the AI system could be improved using this comparison as an indicator of the difference between the AI interpretation and the intended meaning of the user. In addition, user data could be used to customize or train AI models customized for specific use cases, depending on the needs of the customer. From ethical, legal and privacy standpoints, such features would raise vastly different considerations than the current implementation of the prototype. These considerations are also discussed in some of the following sections.

Source

Relevant datasets for the project were identified and discussed. The discussion covered the origin, quality, and expected lifetime of the data, as well as the relevance of the collected information for the project's purpose. The identified data can be divided into two categories: User data and AI training data.

User data in the Smart Care application included user speech input, and form data submitted to ePDoc.net. The assessment considered the quality of these datasets. However, for user data, it was concluded that objective quality assessment was neither possible nor necessary. It was explained that in the current iteration of the project, the user's speech and its transcriptions are processed only in memory of the Smart Care application and the AI system. Submitted form data is saved to the ePDoc.net database and does not expire until 20 years after the client has died.

However, if user data were to be used for AI model training in the future, the situation would change. Speech input transcription and form data would need to be saved in order for a comparison to be made. Saving personal data would require anonymization and a data protection agreement to be in place.

Several factors that may affect the lifetime and relevance of the data were discussed. Legal retention limits determine how long personal data can be stored. However, it was also noted that once data is anonymized and can no longer be linked to an individual, it is no longer subject to data protection laws. In addition, technical changes to models, changes in documentation standards, and the addition of new language support may result in user data or models losing relevance.

During the discussion, it was clarified that the AI system employs a variety of models, primarily general-purpose ones, selected based on the specific task. For example, a small language model might be used for straightforward tasks such as extracting vital parameters from transcriptions, while more complex tasks may require a large language model. The AI team actively evaluates and selects models that best align with system requirements. Because these models are general-purpose and vary by task, the exact datasets used for their training are not always known. However, since the project focuses on documentation rather than clinical decision-making or diagnosis, verifying the origin and quality of the training data was not considered obligatory. In potentially critical use cases, the use of AI models could be restricted to ones that have been pre-approved.

Concerns regarding quality assurance were raised, specifically about ensuring the quality of AI results after modifying models. In response, it was clarified that a release process is in place for managing model changes. Additionally, it was noted that different models can be tested in the initial phases of the project and that the selection can be finalized at a later stage.

Finally, it was agreed that the right information is collected for the purpose of the

project. Both user speech input and form data are crucial for the function of the application. It was noted that the AI system can ignore irrelevant information that might be included in the user's speech, and that AI output includes only content relevant to the context of each module.

Anonymization

For the use case of the Smart Care prototype, anonymization or pseudonymization was not considered necessary. The user's voice input is only processed in system memory, and the form data saved to ePDoc.net must remain identifiable for care personnel. However, anonymization was deemed necessary in future scenarios involving broader data use or extended functionality.

As discussed in earlier topics, the use of user data to improve the AI solution is a possible future use case. In such cases, the data would need to be handled appropriately to meet legal requirements. This would include anonymization, such as removing names or other personal information from the data. During the workshop, it was noted that it was not immediately clear how or where anonymization would be carried out, for example, whether it would be done in the Smart Care application, by the AI provider, or elsewhere. These questions were not answered, but it was noted that systems capable of anonymizing such data already exist.

Visualization

The visualization of the project was not considered highly important in the context of this assessment. The results of the project are presented through the Smart Care application. An alternative way to present the project could be through a native mobile application instead of a web-based one.

Access

Form data sent by the user is stored in Vitec Raisoft data centers in Switzerland. Access to the data is limited to named Raisoft SaaS personnel, certain BESA QSys support staff, and the organizations that use the data. Access is monitored through audit logging and access control. If user data were to be saved by the AI provider, access would be limited to those who need it for their work.

Open access and the re-use of data

As discussed earlier, the use of user data could be considered in the future, provided that it is handled in compliance with data protection law. An initial idea was also

mentioned, which involved asking customer organizations for permission to use their ePDoc data to train customized AI models in exchange for lower licensing fees. However, this was only a preliminary idea, and no further steps have been taken.

7.3 General considerations

Responsibility

Policies, guidelines, and laws relevant to the project were discussed. Data protection law was highlighted as the most important requirement to follow. The GDPR must also be followed, as Vitec Raisoft is an EU-based company. ISO certifications were also discussed. Raisoft is certified according to ISO/IEC 27001:2022 (information security management) and ISO 9001:2015 (quality management). BESA QSys is in the process of acquiring ISO/IEC 27001:2022 certification. Adamcares is not yet ISO certified, but plans to be in the near future. The cloud infrastructure provider used by Adamcares is ISO/IEC 27001:2022 certified. The detailed descriptions of ISO standards were emphasized to help improve internal processes. These certifications were also noted as helpful in building trust with customers.

For the prototype project, responsibility was evenly shared between the parties. BESA QSys acts as the primary data handler for this project and will continue to do so in the future. Raisoft and Adamcares will be listed as subcontractors and must apply the same data use rules as the primary data handler.

Communication

For the Smart Care prototype, no formal communication strategy was established. Communication between the collaborative partners took place through regular and ad hoc meetings, summaries, and business analyses. No marketing-related communication strategies existed for the prototype. It was noted that if the project is continued and developed into a marketable product, more concrete communication strategies and channels will need to be defined.

Transparency

The participants agreed that there is not a high risk of the project creating public concern or outrage. Transparency about the project was supported by conducting a general customer survey on the use of AI. At this stage of the project, the aim was to be as transparent as possible and gauge the interest of potential customers. In addition, the prototype can be demonstrated to interested potential customers. It was

also mentioned that in the possible pilot phase of the project, customers would fill in weekly forms in order to give feedback on the AI output.

The assessment considered whether individuals have the opportunity to raise objections to the results of the project and to opt out of it. The use of the resulting product was concluded to be entirely voluntary, and only those who wish to use it are expected to do so. Smart Care is not a standard feature of ePDoc, but rather an optional add-on.

Privacy

Participants from BESA QSys and Adamcares also serve as data protection officers within their respective organizations. At Raisoft, the data protection officer was not involved during the prototype phase but is expected to participate in the future. When the project progresses beyond the prototype stage, a data protection impact assessment will be done.

Bias

In general, the mood towards the project was enthusiastic, positive and hopeful. The integration of AI in this and other projects was seen as a key driver of these attitudes. In addition, the AI provider highlighted interest in understanding the feedback provided by customers, how it differs from other use cases, and how the original solution can be improved to meet new customer needs.

It was concluded that the prototype project does not pose a risk of contributing to discrimination against individuals or groups, nor does it contain any relevant or harmful bias. This is unlikely to change in the future, at least within the scope of the intended use case.

The prototype implementation does not include a feedback loop. However, as discussed in several earlier topics, user data could potentially be used to improve the solution in the future, thereby establishing a feedback loop. Any such implementation should aim to yield clearly positive results and enhance the system, provided that the data is handled appropriately.

Future scenarios

The long-term effects of this project are expected to differ significantly from the short-term impact of the prototype. This project opens the door to greater AI involvement in improving existing practices and improving the quality of work and life for nurses and care institutions.

The lessons learned from this project can serve multiple purposes, such as analyzing assessment data, improving AI models, and integrating AI-generated documentation directly into the system, without the need for human verification. Potential misuse is difficult to identify as long as data-related considerations are properly managed.

7.4 DEDA assessment conclusion

It can be concluded that organizational and personal values are adequately safeguarded. No contradictions were identified between the listed values and the Smart Care project. The primary objective of the project is to improve the quality of care and improve medical documentation. Trust and transparency are key values of the project, and have been supported through a customer survey on the use of AI, clear high-level explanations of system functionality, highlighting potential benefits, and offering the application as an optional feature.

During the assessment, no major ethical bottlenecks were identified. However, it was acknowledged that some customers may have concerns about the use of AI in medical contexts. Therefore, clear and transparent communication about how the system functions and its limitations is necessary.

The assessment revealed several new insights about the project. The AI system proved to be highly flexible and customizable, adapting to different customer needs. In addition, the use of various models allows the system to balance output quality and response time, for tasks of varying complexity. The prototype also served as an initial step in the integration of AI tools into ePDoc.net and lays the groundwork for future implementation. Finally, it was found that the use of AI for simple input/output operations is relatively straightforward when no personal data is stored. However, the use of personal data for feedback loops, model improvement, and analysis is much more complicated from a legal and ethical perspective.

The Smart Care prototype project has been completed. Further development will begin if there is sufficient customer interest in the product. At this time, there are no ethical concerns that prevent the continuation of the project.

8 Discussion

In this chapter, the results of the DEDA workshop are discussed and compared with the findings of the literature review. The aim is to identify key similarities and differences between the literature and the insights of the stakeholders. This comparison enables a critical reflection on how industry-proposed best practices and ethical considerations align with a real-world implementation. Furthermore, the literature review provides an opportunity to further evaluate the Smart Care prototype and to identify future considerations and potential challenges for further development and deployment in clinical settings.

The importance of human control over AI systems was emphasized in both the literature [31, 34] and the DEDA assessment. The literature highlighted that review and oversight of AI-generated content are essential to ensure accountable use and to preserve trust in AI-powered systems. The design of the Smart Care prototype aligns with these principles by aiming to complement existing workflows rather than replace them. Furthermore, the user retains full control over the documentation process, with the ability to accept, modify, or reject AI-generated content. However, the assessment did not include a detailed discussion on how users might perceive or evaluate AI suggestions in practice. User attitudes and expectations toward AI performance can vary significantly between individuals. In the assessment, it was noted that it would be useful to be able to explain the functionality of the system on a high level, for example, to customers who wish to know more about the system. In addition to this high-level understanding, there should also be sufficient understanding of the limitations of the system, which can be forwarded to end users through onboarding and interactive discussion. Support and education of end users would help in avoiding over-reliance or blind trust, especially if the system is perceived as highly accurate. Critical examination of AI output should be encouraged. Over time, increased awareness and continued experience with the system allows for better understanding of its capabilities, facilitating efficient and responsible use.

In general, it can be said that AI models that have been explicitly trained for a specific purpose with sufficiently broad and context specific data should have a better chance of producing satisfactory results. This also applies to the healthcare sector, which has a lot of medical terminology and abbreviations, and AI models trained with a wide range of medical conditions and scenarios would be needed for best performance [21]. However, this produces the dilemma between practicality and performance. The creation of AI models requires large amounts of data and other resources, and the number of use cases is practically unlimited in the medical field. Therefore, the notion of healthcare organizations only using tailored AI models or even creating them themselves is often not realistic, especially for small institutions. When it comes to producing and modifying text content, general purpose models can be quite effective. However, they still pose the threat of misinterpretation scenarios or words, which can lead to faulty documentation and further care errors.

As revealed in the assessment, the AI service provided by Adamcares uses a multitude of different AI models, which form a pipeline that produces the desired result in the specified format. In addition, the usage of different models can be experimented with. As the service mostly employs general-purpose models, the quality of the datasets that have been used to train these models is often not known. Therefore, it is difficult to evaluate the performance of these models without testing them. The current iteration of the service focuses on the practicality of the implementation, by making use of existing models and adapting to different use cases by using various techniques such as prompt engineering and retrieval augmented generation. This raises questions on whether the output quality of the pipeline is sufficient for handling medical documentation and if there are guarantees on the output since the training data is unknown. However, as a counterpoint, even if AI models were trained and tailored just for the use case of the Smart Care project, it could not be guaranteed that they would always work reliably and never make mistakes. Nevertheless, if human accountability over the produced results is properly enforced, in theory, mistakes made by the AI should never end up in the final content. In that case, the performance of the system becomes a factor on usability rather than output accuracy, as the system working unexpectedly produces additional workload to users.

In the DEDA assessment, the customization and training of AI models using user data was one of the main points of discussion regarding the future of the Smart Care project. This raises a multitude of questions. How and where is data anonymized? How is consent for using the data acquired, from organizations and individuals? How is the data stored and could there be vulnerabilities to attacks such as re-identification or reconstruction? Are trained models robust against adversarial and model inversion attacks? Is the privacy of individual protected by privacy preserving techniques? As we can conclude from the literature review, there is no silver bullet for addressing all of these concerns in a general manner, and they will need to be carefully reviewed for each project. These questions will need to be carefully evaluated if user data will be used for AI training in the future.

The topic of consent was only partially addressed in the DEDA assessment. In particular, the assessment did not explore how consent for data processing is currently managed through the existing EHR system. The GDPR provides clear instructions for obtaining consent from individuals for data processing [26]; therefore, it can reasonably be assumed that such consent is obtained, likely through explicit consent [Art. 9(2)(a)] or as a necessity for medical care [Art. 9(2)(h)]. With the involvement of an additional third-party data processor (AI provider), contracts outlining the terms of processing are required [Art. 28]. Since the Smart Care prototype is an additional feature of existing EHR software, the purpose of data processing remains unchanged (documenting care), but the method changes with the introduction of AI. A definitive legal interpretation would be needed to determine whether this new way of processing falls under the current acquisition of consent, but this is outside the scope of this thesis. However, from an ethical point of view, the implications of this change should not be overlooked, especially since some patients have indicated the desire to opt out

of AI tools in their care [31]. As noted in the assessment, Smart Care is offered as an optional service, which means that on the organizational level the inclusion of this service in the EHR system is voluntary. However, patients should be clearly informed how AI is used and for what purpose. Providing patients with the opportunity to understand, question, and reject AI use in their care process aids in acquiring informed consent and building trust.

In the literature review, the importance of actively involving a diverse set of stakeholders was emphasized when discussing the use of AI [6] and the implementation of AI solutions in healthcare [27]. When it comes to the Smart Care prototype, several stakeholders have already been involved in the early stages of the project. The collaborating organizations that took part in the DEDA assessment each represent their own perspectives when it comes to creating and discussing a new solution. Vitec Raisoft offers perspective on the implementation of the solution and the integration of the AI service in the existing system. Adamcares acts as an expert resource for the AI service and is able to share the required knowledge with other stakeholders. Of the stakeholder organizations, BESA QSys offers the closest perspective to end users and patients. In the future, further involvement of end-users and patients could prove beneficial to fully assess the potential of the solution. This would also help to keep the solution patient/user-centered and facilitate focusing on real care needs.

As mentioned, poor integration of unfamiliar technologies into existing workflows can be a key obstacle to adopting AI in healthcare [20]. Due to the excitement and hype surrounding AI, there is an inherent market pressure to offer AI-based solutions in existing systems. These solutions are exposed to the potential pitfall of creating new tools for the sake of creating new tools, without fully considering the consequences on existing working practices or the usefulness of the proposed solutions. Although this market pressure is a factor for creating the Smart Care prototype, there are clearly other significant factors as well, such as making documenting more efficient and better quality, and facilitating fluent workflow through an easy to use mobile application that requires minimal interaction. In addition, one of the values listed in the DEDA workshop was purposeful innovation, which further aligns the project with the ideology of providing solutions with positive impact.

There are some limitations with respect to the results presented in this thesis. As discussed previously, few studies provide long-term evaluations of human-AI collaboration systems in real-world settings [34]. This thesis faces a similar limitation, as the evaluated system is only a prototype and not yet deployed for use. As a result, the evaluation is based primarily on theoretical analysis and is restricted by the scope of the prototype. Future research should focus on the long-term development and assessment of the tool in real-world use cases to enable for more comprehensive evaluation. In addition, the author of this thesis worked as a software developer for Vitec Raisoft and actively contributed to the implementation of the Smart Care prototype. Although the evaluation and comparison to literature were carried out as objectively as possible, there may be unintentional bias due to the direct participation

of the author in the development of the project.

9 Conclusion

The main objective of this thesis was to outline key considerations for responsible use of AI in a healthcare documentation setting. The literature review covered the most common uses of AI in healthcare documentation, and current challenges regarding privacy and ethics. These findings were complemented by a Data Ethics Decision Aid (DEDA) workshop conducted with industry stakeholders to evaluate the implications of implementing a real-world AI assisted documentation solution.

The growing role of third party companies offering AI solutions for healthcare organizations warrants that the privacy and data of individuals is adequately protected. Organizational and contractual safeguards are essential for the lawful processing of data, and the importance of legal agreements and regulations cannot be understated. In addition, the provided services and the underlying models should be robust against different types of attacks. The aforementioned factors are increasingly important if sensitive health data is used to train domain specific AI models.

The importance of accountability, human oversight, and the use of AI as a tool, rather than a replacement for existing practices were some of the key findings in ensuring responsible use. These measures help mitigate some known AI issues, such as hallucinations or bias, by ensuring that humans stay in control. However, this poses the risk of generating additional workload in an already busy environment, thus a balance between trust and skepticism toward AI is essential for successful integration into existing workflows. This requires that users are properly educated on the potential of AI solutions and their capabilities. It was also found that maintaining transparency on AI use in care is important for eliciting trust between patients, staff, and new technologies. Patients should be aware of how their care is affected by AI tools and have the ability to make an informed decision to opt out. Patients should also be made aware of the potential benefits of AI tools in their care, such as increased patient interaction or improved efficiency of existing processes.

Future research should focus on gathering more data about the use of AI powered documentation tools in practice. This would help the industry design and implement patient centered solutions that bring value to users and patients and ultimately promote high-quality care.

References

- [1] Y. Pinevich et al. “Interaction time with electronic health records: a systematic review”. In: *Applied clinical informatics* 12.04 (2021), pp. 788–799.
- [2] C. Sinsky et al. “Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties”. In: *Annals of internal medicine* 165.11 (2016), pp. 753–760.
- [3] N. Khalid et al. “Privacy-preserving artificial intelligence in healthcare: Techniques and applications”. In: *Computers in Biology and Medicine* 158 (2023), p. 106848.
- [4] B. Murdoch. “Privacy and artificial intelligence: challenges for protecting health information in a new era”. In: *BMC medical ethics* 22.1 (2021), p. 122.
- [5] B. Meskó and E. J. Topol. “The imperative for regulatory oversight of large language models (or generative AI) in healthcare”. In: *NPJ digital medicine* 6.1 (2023), p. 120.
- [6] S. M. Williamson and V. Prybutok. “Balancing privacy and progress: a review of privacy challenges, systemic oversight, and patient perceptions in AI-driven healthcare”. In: *Applied Sciences* 14.2 (2024), p. 675.
- [7] K. Rasheed et al. “Explainable, trustworthy, and ethical machine learning for healthcare: A survey”. In: *Computers in Biology and Medicine* 149 (2022), p. 106043.
- [8] Utrecht University. *DEDA Remote | Data Ethics Decision Aid (DEDA)*. <https://deda.dataschool.nl/en/remote/>. Accessed: 2025-05-19. 2025.
- [9] B. G. Arndt et al. “Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations”. In: *The Annals of Family Medicine* 15.5 (2017), pp. 419–426.
- [10] S. V. Blackley et al. “Physician use of speech recognition versus typing in clinical documentation: A controlled observational study”. In: *International Journal of Medical Informatics* 141 (2020), p. 104178.
- [11] J. Zhang et al. “Intelligent speech technologies for transcription, disease diagnosis, and medical equipment interactive control in smart hospitals: A review”. In: *Computers in Biology and Medicine* 153 (2023), p. 106517.
- [12] J. Joseph et al. “The impact of implementing speech recognition technology on the accuracy and efficiency (time to complete) clinical documentation by nurses: A systematic review”. In: *Journal of clinical nursing* 29.13-14 (2020), pp. 2125–2137.
- [13] S. Latif et al. “Speech technology for healthcare: Opportunities, challenges, and state of the art”. In: *IEEE Reviews in Biomedical Engineering* 14 (2020), pp. 342–356.

- [14] S. V. Blackley et al. “Speech recognition for clinical documentation from 1990 to 2018: a systematic review”. In: *Journal of the american medical informatics association* 26.4 (2019), pp. 324–338.
- [15] J. J. W. Ng et al. “Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: a systematic review”. In: *BMC Medical Informatics and Decision Making* 25.1 (2025), p. 236.
- [16] F. S. Falcetta et al. “Automatic documentation of professional health interactions: A systematic review”. In: *Artificial Intelligence in Medicine* 137 (2023), p. 102487.
- [17] S. A. Mess, A. J. Mackey, and D. E. Yarowsky. “Artificial intelligence scribe and large language model technology in healthcare documentation: advantages, limitations, and recommendations”. In: *Plastic and Reconstructive Surgery–Global Open* 13.1 (2025), e6450.
- [18] D. Khurana et al. “Natural language processing: state of the art, current trends and challenges”. In: *Multimedia tools and applications* 82.3 (2023), pp. 3713–3744.
- [19] R. Bommasani. “On the opportunities and risks of foundation models”. In: *arXiv preprint arXiv:2108.07258* (2021).
- [20] A. A. Tierney et al. “Ambient artificial intelligence scribes to alleviate the burden of clinical documentation”. In: *NEJM Catalyst Innovations in Care Delivery* 5.3 (2024), CAT–23.
- [21] A. Biswas and W. Talukdar. “Intelligent clinical documentation: Harnessing generative AI for patient-centric clinical note generation”. In: *arXiv preprint arXiv:2405.18346* (2024).
- [22] A. J. Thirunavukarasu et al. “Large language models in medicine”. In: *Nature medicine* 29.8 (2023), pp. 1930–1940.
- [23] A. Arora and A. Arora. “The promise of large language models in health care”. In: *The Lancet* 401.10377 (2023), p. 641.
- [24] A. Bracken et al. “Artificial Intelligence (AI)–Powered Documentation Systems in Healthcare: A Systematic Review”. In: *Journal of Medical Systems* 49.1 (2025), p. 28.
- [25] K. Abouelmehdi, A. Beni-Hessane, and H. Khaloufi. “Big healthcare data: preserving security and privacy”. In: *Journal of big data* 5.1 (2018), pp. 1–18.
- [26] European Parliament and Council of the European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>. Accessed: 2025-07-15. OJ L 119, 4.5.2016, p. 1–88, May 4, 2016.
- [27] J. He et al. “The practical implementation of artificial intelligence technologies in medicine”. In: *Nature medicine* 25.1 (2019), pp. 30–36.

- [28] The Federal Assembly of the Swiss Confederation. *Federal Act on Data Protection (FADP) of 25 September 2020 (Status as of 1 April 2025)*. <https://www.fedlex.admin.ch/eli/cc/2022/491/en>. Accessed: 2025-07-25. 2020.
- [29] A. Qayyum et al. “Securing machine learning in the cloud: A systematic review of cloud machine learning security”. In: *Frontiers in big Data* 3 (2020), p. 587139.
- [30] A. Salem et al. “MI-leaks: Model and data independent membership inference attacks and defenses on machine learning models”. In: *arXiv preprint arXiv:1806.01246* (2018).
- [31] J. P. Richardson et al. “Patient apprehensions about the use of artificial intelligence in healthcare”. In: *NPJ digital medicine* 4.1 (2021), p. 140.
- [32] A. Panagopoulos et al. “Incentivizing the sharing of healthcare data in the AI Era”. In: *Computer Law & Security Review* 45 (2022), p. 105670.
- [33] M. A. Ahmad, C. Eckert, and A. Teredesai. “Interpretable machine learning in healthcare”. In: *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*. 2018, pp. 559–560.
- [34] C. Bossen and K. H. Pine. “Batman and Robin in healthcare knowledge work: Human-AI collaboration by clinical documentation integrity specialists”. In: *ACM Transactions on Computer-Human Interaction* 30.2 (2023), pp. 1–29.
- [35] S. K. Bell et al. “Frequency and types of patient-reported errors in electronic health record ambulatory care notes”. In: *JAMA network open* 3.6 (2020), e205867–e205867.
- [36] A. S. Franzke, I. Muis, and M. T. Schäfer. “Data Ethics Decision Aid (DEDA): a dialogical framework for ethical inquiry of AI and data projects in the Netherlands”. In: *Ethics and Information Technology* 23.3 (2021), pp. 551–567.
- [37] B. Latour and S. Woolgar. “The social construction of scientific facts”. In: *Beverly Hills ua* (1979).
- [38] B. Latour and P. Weibel. “Making things public: Atmospheres of democracy”. In: (2005).
- [39] A. McIntyre. *Participatory action research*. Sage publications, 2007.
- [40] F. Baum, C. MacDougall, and D. Smith. “Participatory action research”. In: *Journal of epidemiology and community health* 60.10 (2006), p. 854.
- [41] L. Floridi and M. Taddeo. *What is data ethics?* 2016.
- [42] B. S. Bloom et al. *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. Longman New York, 1956.
- [43] R. K. Yin. *Case study research: Design and methods*. Vol. 5. sage, 2009.
- [44] Vitec Raisoft. *Vitec Raisoft | interRAI Solutions for Healthcare and Social services*. <https://www.vitec-raisoft.com/>. Accessed: 2025-06-03.

- [45] BESA QSys. *Pflegedokumentation ePDoc* | BESA QSys. <https://besaqsys.ch/de/pflegedokumentation-epdoc>. Accessed: 2025-06-05.
- [46] BESA QSys. *Home* | BESA QSys. <https://besaqsys.ch/de>. Accessed: 2025-07-11.
- [47] Adamcares. *Zeit sparen bei der Pflegedokumentation* | Adamcares. <https://www.adamcares.ai/>. Accessed: 2025-07-11.

A DEDA assessment questions

This appendix contains the full list of questions used in the DEDA assessment workshop. The questions follow the structure of DEDA REMOTE V2 [8].

Phase One

1. Project name, date
2. Project team
3. What is the project about and what is its goal?
4. What are possible measures or actions that will be deployed based on the results of this project?
5. What kind of data will you be using? Give a brief description of its contents.
6. Who are the stakeholders of this project and whom/what does it impact?
7. What are the pursued benefits of this project?
8. What problems or concerns could arise in connection with this project?
9. Which of the topics below (algorithms, source, anonymization etc.) are applicable to your project? Decide which might be skipped during the workshop.

Phase Two

This phase can be completed in advance by the project leader, or during the workshop by the project team. Discuss the three most important values of the organization and write them down. Write down a personal value for each participant.

Phase Three

Walk through the questions together and answer them. Create an action item for each question that cannot be answered immediately, and note which values are affected.

Data-Related Considerations

Algorithms

10. Does this project use an algorithm? If not, go to 'Source'.
11. How do you deal with false positives and false negatives?
12. Can someone on the team explain how the algorithm in question works? Is it necessary for someone to be able to explain what it does?
13. Do you consider the outcomes of the model as leading or additional to your decision model?
14. Is there human control over mistakes the algorithm can make? How much room does a person have to deviate from the system when needed?

Source

15. Where does the data come from?
16. Have you checked the quality of the data set(s)?
17. Do the data have an "expiration date"?
18. Are you collecting the right information for your purpose?

Anonymization

19. Is it necessary to anonymize, pseudonymize or generalize the dataset(s)?
20. In the case of anonymization: has it been checked if the data is not traceable?
21. In the case of pseudonymization: who has the key to reverse the pseudonymization?

Visualization

22. How will the results of the project be presented? Will the results be visualized?
23. What alternative ways of visualizing the results are possible?

Access

24. Who has access to the data set(s)?
25. How is access monitored?

Open Access and the Re-Use of Data

26. Are (parts of) the (input) data suitable for re-use? If so, under which conditions?
27. Are the results suitable for re-use? If so, under which conditions?

General Considerations

Responsibility

28. Are there policies or guidelines within the organization that apply to this project?
If so, which ones?
29. Who is/are ultimately responsible for the project?
30. Are the tasks and responsibilities of that person/those people clear, with regard to this project?
31. Is this project suitable for collaboration with (commercial) partners? If so, which parties might those be?

Communication

32. What is the communication strategy for this project (both for its positive as well as negative impact)? In the case of collaborative partners: has this strategy been coordinated with them?
33. What communication strategies have been defined in case something fails?
34. Who is responsible for creating these strategies?

Transparency

35. Does the project risk generating public concern or outrage, now or in the future?
36. How transparent are you about this project towards citizens?
37. How are citizens actively involved?
38. Do citizens have the opportunity to raise objections to the results of the project?
39. Can citizens opt out of their involvement in the project? If so, when and how?

Privacy

40. What laws, regulations and/or guidelines apply to your project?
41. Did you involve the privacy officer and/or data protection officer in the project?
42. Have you carried out a data protection impact assessment (DPIA)?

Bias

43. What is your gut feeling about this project? Do you have any concerns?
44. Is there a risk that the project could contribute to discrimination against certain people or groups?
45. Are all relevant citizens adequately represented within your data? Which citizens are missing or under-represented?
46. Does your model have a feedback loop that could create negative consequences?

Future Scenarios

47. Function creep: can you imagine a future scenario in which the outcomes of this project are (mis)used for another purpose?
48. Do your answers to these questions change when you consider possible long-term effects? Why?
49. When will this project be evaluated?

Phase Four

In this final step of DEDA REMOTE, the project team forms a conclusion. Together, discuss and answer concluding questions based on your answers to the data-related and general considerations.

1. Are organizational values and personal values adequately safeguarded?
2. What are the main ethical bottlenecks?
3. What are new and surprising insights?
4. Under what conditions are we willing/not willing to proceed with this project?
5. Review all the action items listed below and write down the action holder for each one.