Evaluation of Ambient AI-Based Clinical Documentation in Physiotherapy

A case study of Western Uusimaa Wellbeing Services County

John Sundberg

Supervisor:

Jyri Rajamäki, Principal Lecturer, Laurea University of Applied Sciences

Date: 28.2.2026

Co-funded by the European Union

Evaluation of Ambient AI-Based Clinical Documentation in Physiotherapy

A case study of Western Uusimaa Wellbeing Services County

John Sundberg

Supervisor:

Jyri Rajamäki, Principal Lecturer, Laurea University of Applied Sciences

Date: 28.2.2026

Co-funded by the European Union

Acknowledgements

I would like to thank principal lecturer Jyri Rajamäki for his guidance and supervision throughout the thesis process.

I also wish to thank Sarita Aho, Sanna Vuorinen and Tuuli Pajunen at LUVN for supporting the implementation and facilitating the conditions that made this research possible.

I am grateful to all the physiotherapists at LUVN physiotherapy services who volunteered to participate in the different aspects of this research and engaged actively in the pilot project.

I thank Gosta Labs and Juuso Parkkinen for facilitating access to- and support with processing of the raw data.

Finally, I would like to thank Filip, Sarita and my parents for proofreading the manuscript and providing valuable feedback on language and clarity.

Abstract

Ambient AI- assisted documentation shows significant potential for reducing documentation workload in healthcare. Empirical evidence on AI-assisted documentation in a physiotherapy setting remains limited, particularly outside English-language contexts. This case study evaluates the implementation of a Large Language Model (LLM)-based ambient AI scribe in public outpatient physiotherapy services at Western Uusimaa Wellbeing Services County in Finland.

A longitudinal single-case design was applied using multiple data sources, including automatically logged timestamp data, user ratings, bi-weekly questionnaires and a self-reported control group comparison. Documentation time was analysed using clinician-level medians and non-parametric statistical comparisons. Sensitivity analyses were performed to address potential artefacts in the timestamp data.

Median documentation time was lower in the AI scribe group compared to the control group, with reductions ranging from 27 % to 39 % depending on timestamp processing assumptions. Considerable between-clinician variability was observed, suggesting that time savings depend on workflow adaptation and documentation practices. User rating and error rate improvements were particularly related to an update to the speech recognition model. Low to somewhat positive effects were reported on therapeutic alliance and job satisfaction, though survey response rates were limited.

The findings suggest that ambient AI-assisted documentation is viable in public outpatient physiotherapy in a Nordic language setting and may reduce documentation burden. The results contribute empirical evidence from a physiotherapy setting and highlight the importance of workflow integration to realize the full potential of AI-assisted documentation.

Keywords: Ambient AI | AI Scribe | Clinical documentation | Physiotherapy | Documentation efficiency

JEL Classification: I11, O33

Index

Acknowledgements ii

Abstract iv

Index vii

Glossary ix

Introduction 1

Context 1

The Documentation Burden 1

Documentation in Physiotherapy 2

Problem Description 3

Aim and Objectives 4

Scope and Delimitations 4

Relevance 4

Contribution 5

Structure of the Thesis 5

Use of AI 5

1. Background 6

1.1. Speech Recognition 6

1.2. From Speech Recognition to Digital Scribes 6

1.3. The Era of LLMs and Ambient AI 7

1.3.1. Findings of Systematic Reviews Looking at Ambient AI Scribes 7

1.3.2. Studies on Rehabilitation 8

1.3.3. First Large-Scale RCT 9

1.4. Synthesis of Current Evidence 9

2. Methodology 11

2.1. Research Design 11

2.2. Case Description 11
2.3. Unit of Analysis 12
2.4. Data Sources and Triangulation 12
2.5. Data Collection 13
2.6. Data Analysis 14
2.7. Ethical Considerations 15
2.8. Reflexivity and Researcher Role 15
2.9. Model Updates and Other Interventions 16
3. Results 17
3.1. Time Stamp Data 17
3.1.1. Relative Variability of Documentation Time 18
3.1.2. Temporal Evolution of Documentation Time Distribution 19
3.1.3. Time Stamp Data from the Control Group Survey Period 19
3.2. Control Group Data 21
3.3. Time Spent on Documentation, AI Scribe vs Control Group 22
3.4. User Feedback 23
3.4.1. User Ratings 23
3.4.2. Reported Errors 24
3.5. Bi-Weekly User Questionnaire 25
4. Discussion 28
4.1. Time Spent on Documentation 29
4.2. Reported Errors in AI Documentation 31
4.3. Bi-Weekly Questionnaire 32
4.4. Language 32
4.5. Limitations 33
5. Conclusion 34

5.1. Findings 34
5.2. Limitations 34
5.3. Future Research 35
6. Bibliographic References and Sources 37
7. Annexes 39
7.1. Annex A – Literature Search Strategy 39
7.1.1. Search Strings 39
7.1.2. Screening and Selection Process 40
7.1.3. Inclusion Criteria 40
7.1.4. Exclusion Criteria 40
7.1.5. Results of the Search 41
7.2. Annex B – Research Plan 41
7.2.1. Working title 41
7.2.2. Relevance 41
7.2.3. Problem Definition and Research Gap 41
7.2.4. Aim and objectives 42
7.2.5. Data collection and Analysis 42
7.2.6. Scope and Limitations 43
7.2.7. Ethical considerations 44
7.2.8. Research group 45
7.2.9. Key references 45
7.3. Annex C – Informed Consent Form 45

Index of Figures

Figure 0.1 Documentation headings and their contents 2
Figure 3.1: Median documentation time per month, in seconds 18
Figure 3.2: Relative variability over time 18
Figure 3.3: Temporal evolution of documentation time distribution, in seconds 19
Figure 3.4: Box plot of AI scribe group at 30-minute filtration 20
Figure 3.5: Box plot of AI scribe group at 20-minute filtration 20
Figure 3.6: Box plot of AI scribe group at 10-minute filtration 21
Figure 3.7: Box plot of control group documentation time 21
Figure 3.8: Average user rating over time 23
Figure 3.9: Average user rating over time, most active users 24
Figure 3.10: Reported errors by type over time 25
Figure 3.11: Bi-weekly questionnaire Q1 26
Figure 3.12: Bi-weekly questionnaire Q2 26
Figure 3.13: Bi-weekly questionnaire Q3 27
Figure 3.14: Bi-weekly questionnaire Q4 27

Index of Tables

Table 2.1: Documentation by type 13
Table 2.2: Bi-weekly questionnaire responses 13
Table 3.1: Initial visit statistics, times expressed in seconds 17
Table 3.2: Relative variability of documentation time (%) 18
Table 3.3: Temporal evolution of documentation time distribution, in seconds 19
Table 3.4: AI scribe group documentation statistics at different filtration points 20
Table 3.5: Control group documentation statistics 21
Table 3.6: Documentation time differentials between control group and AI group at different filtration points 22
Table 3.7: Mann-Whitney U test comparisons between control group and AI group at different filtration points 22
Table 7.1 Timeline 44

Glossary

AI scribe – An ambient AI based clinical documentation system that uses automatic speech recognition and large language models to continuously capture clinical encounters and generate structured clinical documentation for the clinician

Ambient AI – A technology that uses sensors to gather data from the environment without requiring active user input

ASR – Automatic Speech Recognition

EHR – Electronic Health Record

ICF – International Classification of Function

Filtration – The process of removing specific data points from raw log files to isolate relevant activity. In this study, filtration refers to the removal of time intervals between text editing events, to eliminate non-documentation related tasks from the calculation of total documentation time.

ICPC-2 – Second edition of the International Classification of Primary Care

LLM – Large Language Model

LUVN – Länsi-Uudenmaan Hyvinvointialue or Western Uusimaa Wellbeing-Services County in English

Onboarding – The training process intended to make a new user able, confident and productive with the AI scribe

SR – Speech Recognition

Therapeutic alliance – As defined by Bordin (1979), cited by Evans et al. (2025) refers to “the collaborative and affective bond between a patient and a therapist”

Time stamp – The time automatically recorded by the AI scribe system for events related to the documentation process

Introduction

Context

The Documentation Burden

A large part of the workload in healthcare is generally related to documentation (Sasseville et al., 2025; Bracken et al., 2025). The outpatient setting has a larger influx of new patients relative to the complete workload compared to inpatient care. This highlights the significance of documentation, as assessments of patient condition are made that affect how the care continues. Healthcare professionals need to make sure that documentation includes all the relevant information of the patients' status, for the safety of both the patient and the professional.

Clinical documentation in Finland is guided by multiple layers. Starting from the legislative level, the Ministry of Social Affairs and Health Decree on Patient Records (94/2022) specifies that the health record needs to include the following information: reason for the visit, anamnesis, current status, observations, examination, problems, diagnosis, conclusions, care plan, -implementation and -follow-up, course of the illness and discharge summary.

The Finnish Institute for Health and Welfare (THL, 2025) has created a set of guidelines for how to structure health record texts. The THL guidelines elaborate on the requirements defined by the Ministry, providing a clearer structure for healthcare professionals and providers to follow. Figure 1 shows a summary on the recommended structure. Larger organizations such as the Wellbeing Services Counties additionally have their own guidelines for the professionals to follow, striving to unify the contents and layout of healthcare documentation within the organization.

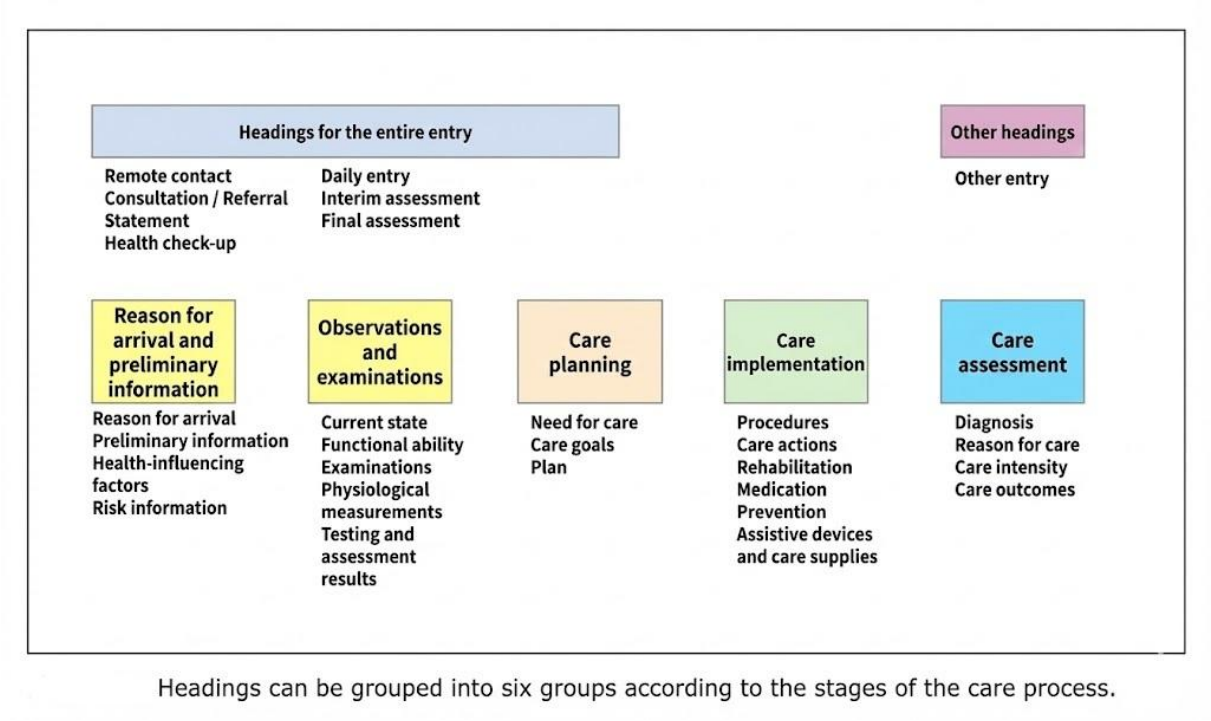


Figure 0.1 Documentation headings and their contents

Adapted from Potilastiedon kirjaamisen yleisopas v 7.0, Terveyden ja Hyvinvoinnin Laitos (THL). 2025. Adaptation using Google Nano Banana.

Documentation in Physiotherapy

Evaluation of functional ability and physical status are at the centre of the initial visit in outpatient physiotherapy, usually providing a functional diagnosis and serving as the base on which the rehabilitation is built. Thus, documentation of this part of the visit, which often contains parts which may not be fully verbalized, is in heavier focus compared to most other healthcare professionals. While a few different frameworks are used to define the structure of the physiotherapy texts [Fysioterapianimikkeistö, International Classification of Function (ICF), International Classification of Primary Care (ICPC-2)], at this time, the Electronic Health Record (EHR) systems often put limitations on integration between AI scribes and the EHR (Leung et al., 2025), so the value of producing these codes for the documentation is yet to be realized.

Physiotherapy documentation in LUVN, following the THL guidelines shown in Figure 1, contains the following headings and contents:

- Reason for arrival
- Patient history

    - Contains preliminary information, health-influencing factors and risk information.
- Assistive devices
- Functional status
    - Contains the current state, functional ability, examinations, physiological measurements and a functional diagnosis.
- Testing and assessment results
    - Contains standardized test results
- Care goals
- Rehabilitation
    - Contains any therapy and exercise performed during the appointment, along with exercise prescription and guidance.
- Plan
    - Contains the physiotherapy plan, consultations and referrals.

Problem Description

Physiotherapists still spend a significant part of their workday on clinical documentation (Evans et al., 2025). In the current healthcare environment, with high pressure on efficiency, ambient AI scribes bring the promise of both reduced time spent and cognitive burden on healthcare professionals (Sasseville et al., 2025; Bracken et al., 2025).

Currently, most of what we know about ambient AI scribes comes from a physician outpatient setting from the United States (Leung et al., 2025). While outpatient physical therapy shares some of the structure for documentation, there is usually a heavier focus on movement and function in the form of observations and clinical tests. This may challenge voice-to-text models as more of the relevant information might only be implied, not directly voiced out loud (van Buchem et al., 2021).

There is very limited research about AI scribes in a physiotherapy setting, and no research on how ambient AI scribes work for clinical documentation in the Nordic languages. Thus, there is a need to assess how ambient AI documentation tools function in a physiotherapy setting, mostly using a less well-known language.

The primary research question of the thesis is how an AI scribe affects the temporal burden of documentation in public healthcare outpatient physiotherapy setting.

Aim and Objectives

The aim of this thesis is to evaluate how the implementation of ambient AI-assisted clinical documentation in physiotherapy services at LUVN affects documentation time and user experience of physiotherapists, using a Case Study Research approach.

Objectives:

1. To assess how ambient AI affects time spent on documentation
2. To evaluate how ambient AI affects physiotherapist job satisfaction
3. To evaluate how ambient AI affects patient interaction
4. To generate insight into what kind of technical and workflow-related improvements lead to better quality of documentation

Scope and Delimitations

This thesis focuses on evaluating the implementation of one specific LLM-based ambient AI scribe in outpatient physiotherapy in the rehabilitation unit of LUVN. Focus will be on workload as most of the quantitative data is focused on this perspective. Information regarding cognitive load and therapeutic alliance are self-reported by the participants of the study. The users of the AI scribe are all volunteers, which may affect the results. Speech is recorded in either Finnish, Swedish or English, and the clinical documentation is mostly in Finnish, with some subgroups documenting in the language used during the visit. The data does not differentiate language per visit, so there is no way to analyze this more specifically. Due to the scope of the Thesis, the data is from a limited timeframe, which weakens the results.

The writer of this Thesis is involved both in the project management of the AI scribe pilot project and as a user of the ambient AI scribe, so there is a risk for bias in any evaluations not relying strictly on quantitative data.

The study does not include analysis of documentation quality or language specific performance analysis. These aspects were considered but excluded due to limitations in data availability and scope.

Relevance

The Ambient AI scribe pilot project at LUVN, in collaboration with a Finnish AI startup, Gosta Labs, started in the spring of 2024 for medical doctors and was expanded to physiotherapy in March 2025, as the first public healthcare outpatient physiotherapy unit in Finland to use an

AI scribe. The pilot project provides a unique opportunity to evaluate how the technology works in a real-world clinical setting.

Contribution

The main contribution of this thesis is to show that an ambient AI scribe can reduce the documentation workload of physiotherapists in a public healthcare outpatient setting outside the English language context. The study additionally shows that potential gains vary significantly based on user workflows and other non-technical aspects of use.

Structure of the Thesis

Chapter 1 – Introduction. Introduces the context, problem description, aim, scope and key concepts.

Chapter 2 – Background. Reviews the literature on the subject, starting with a short recap of the history of speech recognition, with a short review of the current state of research, identifying gaps relevant to this thesis.

Chapter 3 – Methodology. Describes the research design, data sources and analytic methods.

Chapter 4 – Results. Presents the results based on collected data and stakeholder experience.

Chapter 5 – Discussion. Discusses the findings in relation to existing literature, and their implications for practice and policy.

Chapter 6 – Conclusion. Concludes the thesis and outlines recommendations for future research.

Use of AI

AI has been used in this thesis for brainstorming, to support the literature search process and to aid with the chapter structure of the thesis. AI was also used to draft the informed consent form and to adapt or edit some of the figures. AI tools used are ChatGPT, Google Gemini, Microsoft Copilot and Google Nano Banana.

1. Background

While Ambient AI and the digital scribe technology enabled by the modern Large Language Models (LLMs) are new, different forms of automated documentation have been used for decades. Early attempts date as far back as the 1960s, but these systems were usually very specific and limited in function (Turpin, Patricia. 2005). Blackley et.al. (2019) note that early variants of Speech Recognition (SR) were adopted in the late 1980s. This chapter will shortly discuss the history of automated documentation in healthcare. Speech recognition enabled the first versions of automation for narrative clinical documentation and thus will be the starting point for this literature review.

The pace of improvement is rapid. As noted by Sevilla et al. (2022), the scaling of training compute has accelerated rapidly since the advent of Deep Learning in the early 2010s, doubling approximately every 6 months. While not directly showing the causation, this reflects the subjective observation that improvements in the LLMs we use come at a fast pace, potentially leading to big improvements from one generation to the next. While the benefits don't necessarily transfer directly, the impact can clearly be noticed in practice while using the ambient AI models.

1.1. Speech Recognition

The systematic review by Blackley et al. (2019) shows how adoption started from radiology and spread slowly to other specialties. Most systems used dictation into templates, to replace earlier dictation transcribed by typists, rather than fully automatic note generation. While accuracy rates improved over time, error rates varied and outputs required proof-reading and correction.

Hodgson & Coiera (2016) came to similar conclusions. While SR can lead to shorter times to produce a report, an increased correction workload falls on the clinician. SR technology also leads to new safety risks due to clinically important facts sometimes being misinterpreted and documented incorrectly. While SR sped up the transcription process, it did not lead to fundamental improvements to the documentation process. Hodgson & Coiera even note that SR often even led to increased cognitive load for clinicians.

1.2. From Speech Recognition to Digital Scribes

Advances in Natural Language Processing (NLP) led to a shift from dictation towards the idea of a Digital Scribe (DS) in the late 2010s. Coiera et. al. (2018) is one of the early attempts at

imagining a digital scribe combining Automatic Speech Recognition (ASR) with a model that extracts key information to generate structured documentation in the background.

Van Buchem et al (2021) did a scoping review of the pre LLM era of digital scribes, using ASR and NLP to automate clinical documentation. While the results were promising, the outputs were mostly focused on extracting symptoms, classifications or short summaries from the transcript, not fully fledged Electronic Health Records (EHR) documentations.

As recently as 2023, Falcetta et al., in their systematic review, noted that none of the reviewed studies presented a commercially available product. Only one of the 8 included studies presented user evaluation of the prototype, and that model failed the usability test with a 45/100 score, with 68 being the limit for qualifying as being generally usable.

It is also worth noting that none of these early stages of SR, ASR or DS solutions have been researched in a rehabilitation or physiotherapy setting.

1.3. The Era of LLMs and Ambient AI

As in many other industries, the introduction of ChatGPT and the transformer-based LLMs, have led to a new push towards commercially viable digital scribes. As this is still a new technology, terminology is yet to be defined, as pointed out by Leung et al. (2025). The underlying Automatic Speech Recognition (ASR) is a well-established term, but competing terminology such as Ambient Digital Scribing (ADS) and Ambient Documentation Technology (ADT) is used. For this literature review, searches were performed in PubMed and Google Scholar, using ChatGPT 5.1 and Gemini 3.0 as well as references from the relevant articles to search for additional sources. For a detailed description of the literature search strategy, see Annex A. The last search for resources for this thesis was performed 1.12.2025. Five systematic reviews on AI scribes used for clinical documentation were found. None of these discuss physiotherapy or rehabilitation as a specialty. Only two articles were found that included physiotherapy: Evans et al. (2025) and Omon et al. (2025). To illustrate the most recent developments of the field, one Randomized Controlled Trial (RCT), published 26.11.2025 by Lukac et al, was included as well.

1.3.1. Findings of Systematic Reviews Looking at Ambient AI Scribes

Of the 5 included articles, Perkins et al. (2024) included studies ranging from 2005 to 2024. The study didn't find any larger trends, which makes sense as it looks at a very heterogeneous

set of technologies. They note improvements to subsets of the clinical workflow, but no major benefits.

In a similar vein, Ng et al. (2025) included studies from 2001 onwards. As in the Perkins study, the results were varied, with no clear efficiency gains to be found: some systems reduced documentation time and improved completeness of notes, while others even increased editing burden. The authors themselves point out that future studies should be aimed at more complete end-to-end solutions.

Sasseville et al. (2025) included a mix of fully LLM-based and earlier digital scribes. They found that AI scribes positively impact provider engagement, while showing promises of efficiency gains and improvements to clinical workflow. While effects on burnout remain limited, users often reported a reduction in perceived documentation burden. Documentation quality remained a concern.

Alboksmaty et al. (2025) focused on studies from the US, Bangladesh and the Philippines. They found clear gains in effectiveness and efficiency, as well as improvements to patient-provider interaction. Safety concerns remained, due to errors and incomplete information. It is worth noting that not all the included studies looked at full LLM-level ambient scribes.

Bracken et al. (2025) looked only at ambient AI era systems, from late 2023 until august 2024. They found clear improvements to documentation efficiency, with users reporting reduced task load. Regarding quality the results were mixed, with some users expressing concerns about reliability and validity of AI- generated documentation.

Common to all the above-mentioned studies is that they either were small, including between 8 and 11 small, lower quality and mostly quite heterogeneous studies, or that they included aging technologies and an even larger variety of solutions. Only Bracken et al. included purely LLM-based solutions. Nevertheless, while the literature shows a growing evidence base for AI-based clinical documentation, it remains quite limited for now. As mentioned before, these studies primarily focus on the physicians' perspective, leaving physiotherapy in the dark.

1.3.2. Studies on Rehabilitation

As mentioned in Section 2.3, only 2 studies including physiotherapists were found during the literature search. Of these, Omon et al (2025) researched an in-patient setting. Of the already small sample size of 12 participants, only 2 were physiotherapists. The study found no significant change in documentation time, with a decrease in time pressure felt by the

professionals and an increase in documentation quality scores. The study does however highlight the need to teach workflow design and AI literacy to users to reach good results. The study does not discuss specifics about the documentation such as physical status or exercise prescription.

Of the 119 allied health professionals (healthcare professionals other than medical doctors or nurses) participating in the mixed methods study by Evans et al. (2025), 91 were physiotherapists. Participants were recruited from one private allied health organization in Australia. The study found significant reduction to time spent in notes, decreased documentation burden and a positive impact on therapeutic alliance. While the quantitative data doesn't specify how much of the text editing work was done outside the EHR, the qualitative results suggest that there were significant reductions in documentation time.

1.3.3. First Large-Scale RCT

The first large-scale RCT on the subject was published by Lukac et al. (2025), looking at 238 physicians, randomized into 3 groups: two different AI scribe groups and a control. The study looked at two competing AI scribes in the medical setting: Nabla, developed by a US-based startup and Dragon Ambient eXperience Copilot (DAX), developed by Microsoft in partnership with Nuance. Of the two AI scribes, the study found a 9.5% reduction in time spent on notes with Nabla, while with DAX there wasn't a significant difference. Both groups reported a positive impact on burnout, work exhaustion and task load. The quality of the quantitative data is also put to question, as the researchers had not accounted for edits made in the AI scribe platforms. The results also include all documentation during the study, while physicians assigned to the intervention groups only used the tools for 33.5 % (DAX) and 29.5% (Nabla) of their documentation during the study.

1.4. Synthesis of Current Evidence

While there still is a distinct lack of rigorous research on the subject, some trends can be seen. There seems to be a clear positive impact on cognitive load when using AI scribes. While quantitative data is still of low quality, there also seems to be a small to moderate positive impact on documentation time. While physiotherapists and other allied health professionals are still very underrepresented in the literature, the Evans study shows promising results for this group.

It is also worth noting that a vast majority of studies have been made on the English language, not accounting for potential problems caused by speech recognition in smaller languages.

2. Methodology

The analytical framework of this study is grounded in case study research methodology, basing data collection, analysis and interpretation on Yin (2018).

2.1. Research Design

This study applies a single-case study research design, as defined by Yin (2018), to evaluate the implementation of an ambient AI-assisted clinical documentation system in outpatient physiotherapy services at LUVN.

A case study approach was chosen because the phenomenon under investigation – the real-world use of an ambient AI scribe – is context-dependent, evolving over time and cannot be meaningfully separated from its organizational, technical and workflow environment. The study focuses on understanding how the introduction of the AI scribe affects documentation time, job satisfaction, and user experience in a clinical setting.

The design is explanatory and longitudinal, combining retrospective system log data with user feedback collected during the pilot project period.

2.2. Case Description

The case examined is the implementation of an LLM-based ambient AI scribe, developed by the Finnish AI company Gosta Labs, in outpatient physiotherapy units within LUVN. It is a part of a larger ongoing pilot project for the implementation of the AI scribe across LUVN healthcare units. Within LUVN, the AI scribe was first introduced for physicians in the spring of 2024. The project was expanded to physiotherapy in March 2025, making it the first known implementation of its kind in public outpatient physiotherapy in Finland. The documentation template was optimized for receptions with patients presenting musculoskeletal problems, as this is largest patient group in public outpatient physiotherapy. Optimization of the template has been an ongoing process throughout the pilot project, with feedback gathered from participants being used for further improvements.

The AI scribe captures audio during clinical encounters using Automatic Speech Recognition (ASR) and generates a structured draft of the clinical documentation, which the clinician reviews and edits in the web application. Due to technical limitations of the EHR, the text is then manually copy-pasted over into the EHR. Final responsibility for documentation accuracy remains with the clinician.

Participants consisted of licensed physiotherapists working in outpatient rehabilitation services. Participants have been recruited on a voluntary basis, starting with a smaller group from one physiotherapy unit in March 2025 and expanding to all outpatient rehabilitation units in May 2025.

The AI scribe underwent multiple iterative updates during the study period, including several updates to the model structuring the text, and one major update to the speech recognition model. These updates are treated as part of the implementation context rather than as controlled interventions, reflecting real-world deployment conditions. The pilot project also included onboarding sessions and recurring sparring sessions with the clinicians, striving to collect feedback for model development and share advice on how to get the most out of the AI scribe.

2.3. Unit of Analysis

The unit of analysis in this case study is the time spent on clinical documentation. Only initial appointments were measured for the study, as they in general are longer texts and are responsible for the bulk of the time spent on documentation in outpatient physiotherapy.

2.4. Data Sources and Triangulation

Multiple data sources for the unit of analysis were used for methodological triangulation, as this strengthens the construct validity of the case study (Yin, 2018).

The following data sources were used:

1. Automatically logged timestamp data, from which time of initial AI text generation, editing- and copying events were isolated to determine the time used for editing.
2. User ratings of generated documentation as well as optional error reporting within the AI scribe interface
3. Bi-weekly user questionnaires, measuring self-reported perceptions of documentation time, therapeutic alliance and job satisfaction.
4. Control group survey data, measuring self-reported documentation time from voluntary physiotherapists of the same physiotherapy units as the AI scribe participants.

For the control group, the study had to rely on self-reported data as there were no tools available for the researcher to get this as system log data.

2.5. Data Collection

Timestamp data and user feedback from appointments using the AI scribe was collected from 14.3.2025 until 12.12.2025. During this time, a total of 5752 documentations were generated, of which 4215 documentations were using the production model for initial physiotherapy encounters, 1177 documentations using the model for follow-up appointments and 360 documentations using a beta testing model. Table 1 shows a breakdown of documentations by month and type.

Table 2.1: Documentation by type

Month	Follow-up	Initial	Beta	Grand Total
March		91		91
April		72		72
May		194		194
June		480		480
July		368		368
August		458		458
September	65	571	23	659
October	382	707	110	1199
November	464	795	154	1413
December	266	479	73	818
Total	1177	4215	360	5752

The Bi-weekly questionnaire issued by the AI company, was updated mid-September. Included data was collected between 16.9.2025 and 12.12.2025. From the average 31 active weekly users, there were between 3 and 15 responses for each 2-week period. Table 2 shows a more detailed breakdown of responses to the questionnaire.

Table 2.2: Bi-weekly questionnaire responses

Date Range	Respondents
16.9.2025 - 29.9.2025	5
30.9.2025 - 13.10.2025	3
14.10.2025 - 27.10.2025	15
28.10.2025 - 10.11.2025	13
11.11.2025 - 24.11.2025	11
25.11.2025 - 8.12.2025	8
9.12.2025 - 13.12.2025	11
Total	66

Control group data was collected from a two-week period between 17.11.2025 and 28.11.2025. The questionnaire was sent out to all physiotherapists belonging to the outpatient rehabilitation unit in LUVN, requesting all physiotherapists currently not using the AI scribe to respond. The questionnaire got a total of 17 respondents logging their time used for documentation for 229 initial appointments.

2.6. Data Analysis

Quantitative analysis focused on descriptive statistics, temporal trend analysis across the pilot project period and comparison between AI scribe users and the control group.

The time stamp data was first manipulated in Power Query for Excel to filter out the relevant information, and calculations and generation of graphs were done in Microsoft Excel. Graphs containing localized content in Finnish were recreated using Google Nano Banana.

The original timestamp data contained editing events, where a pause of more than 0,5 seconds in activity led to a new event being logged. To address extreme outliers likely caused by factors outside of documentation work, time gaps over 30 minutes between two subsequent events were removed from the total time of a documentation. As physiotherapists at LUVN currently have 15 minutes reserved for each documentation, this cutoff only removes the most extreme outliers, to ensure that the filtration doesn't cause a positive bias to the result.

Similarly, any documentation shorter than 10 seconds from the first edit to the last copying event were excluded from the results under the assumption that the documentation process could not reasonably have been completed in this time. Again, a cutoff that only removes the most extreme outliers, as the researcher timed just the copying process, taking one section at a time, taking about 40 seconds.

Comparisons were made between cutoffs at different gaps between editing events, also looking at removal of time gaps of 10 or 20 minutes between subsequent events. As the temporal trends remained very similar, the initial data was deemed sufficient for deeper analysis for this study.

Due to the non-normal distribution, high variance of documentation times and long tails of the data, likely containing some outliers caused by the logging method, median-based measures were emphasized over means. This is motivated for data containing outliers and long tails that could otherwise skew the results (Conover, 1999). Interquartile range (IQR) was calculated as it captures the spread of the central 50% of observations while remaining

resistant to extreme values. Additionally, relative variability (IQR divided by median) and percentile-based values were examined to analyze distributional spread and tail behavior. These metrics were included to analyze documentation heterogeneity over time in addition to central tendencies.

For group comparisons, median values were calculated at clinician level across included appointments. This reduces the influence of highly active individual users and better reflects between-clinician variability in documentation. These median values were used to calculate the differences between the AI scribe group and the control group, using the non-parametric Mann-Whitney U test applied to the clinician-level medians.

For group comparisons, the data filtrations at time gaps of 10 or 20 minutes made a significant difference in the end results. Thus, the data from all 3 will be provided in the results section to provide deeper insight into the documentation workload.

2.7. Ethical Considerations

A research permit was applied for and approved by the proper authority within LUVN (See Annex B and C for the research plan and informed consent forms provided as part of the permit application). The study was conducted using anonymized and pseudonymized system data. No patient information was used for this study. Participation in questionnaires was voluntary, and no individual user data is reported in a manner that would allow identification. All collected data was stored in LUVNs cloud storage and will be deleted after the thesis is accepted in accordance with LUVN data management policies.

2.8. Reflexivity and Researcher Role

The researcher was involved in project management, model development and as an end user of the system. This is acknowledged as a potential source of bias, especially in the interpretation of qualitative observations. On the other hand, this allows unique insight into the process from multiple perspectives.

To mitigate the risk of bias, the analysis prioritized system-generated quantitative data and descriptive statistical trends over interpretive qualitative claims. Subjective observations from user feedback, sparring sessions and the researcher are reported cautiously and used primarily for context. Limitations related to researcher involvement are further discussed in the Discussion chapter.

2.9. Model Updates and Other Interventions

From the start of the pilot project until the end of the period included in this study, there were 7 major model updates and one update to the speech recognition model. The model updates were implemented 26.3, 11.4, 16.5, 2.6, 24.6, 3.10 and 5.12. The speech recognition model was updated 9.9. The 3.10 update included language support for English and Swedish.

Onboarding of the initial group of 8 testers was 18.3, the same day that use started. An additional 23 users were onboarded 20.5. Voluntary 30-minute sparring sessions were organized in smaller groups 10.6-18.7 on a weekly basis. The format of the sparring sessions was changed to mandatory on a bi-weekly cycle and were held in this format 1.10-16.12.

3. Results

3.1. Time Stamp Data

Table 3 shows a detailed breakdown of key statistics of the initial appointment documentations recorded during the pilot project, on a month-by-month basis.

See section 3.6 for the breakdown on how the time stamp data has been processed.

Table 3.1: Initial visit statistics, times expressed in seconds

Month	Amount	Median	P95	P5	IQR	IQR/Med	Highest	Lowest
March	67	636,48	2055,51	84,66	846,51	133,0 %	3061,23	37,68
April	51	560,44	2258,14	27,14	812,51	145,0 %	3195,89	16,44
May	157	775,29	1891,49	33,60	879,77	113,5 %	3359,13	10,64
June	388	666,02	2145,84	78,40	788,76	118,4 %	3231,43	15,96
July	307	794,44	2263,06	108,19	790,29	99,5 %	3680,38	10,61
August	376	865,92	2571,02	119,42	1030,27	119,0 %	5738,52	12,72
September	470	715,81	2220,98	56,38	906,92	126,7 %	6152,98	13,87
October	626	749,41	2223,90	84,17	842,21	112,4 %	5130,95	10,72
November	708	718,67	2093,41	128,59	610,60	85,0 %	5093,16	13,14
December	424	664,28	1933,07	56,22	667,42	100,5 %	3526,40	15,31

The usage has mainly been growing during the test period. The initial onboarding was done on 18.3.2025 for 8 users at one healthcare station. The pilot project was expanded to include any volunteers in May, with onboarding for an additional 23 users 20.5.2025. After that any volunteers have been allowed to join on request.

The median time spent per documentation was at its highest in august at 14 minutes and 42 seconds, after which the trend has been downwards. It is noteworthy that the median time in March and April was lower than in December. Figure 2 shows the temporal evolution of documentation time medians over the data collection period.

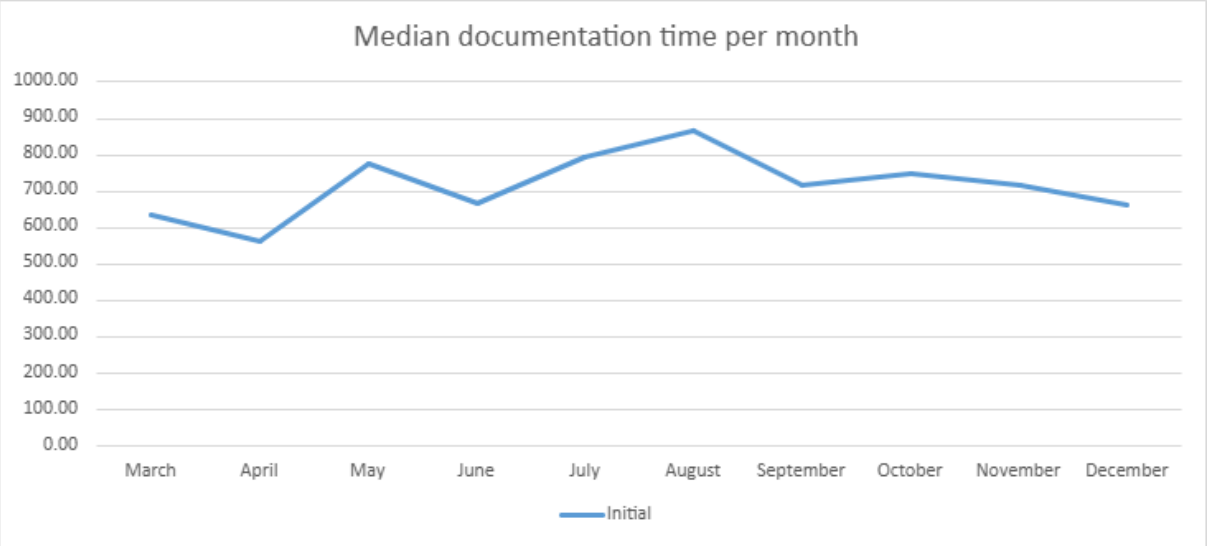


Figure 3.1: Median documentation time per month, in seconds

3.1.1. Relative Variability of Documentation Time

Table 4 shows the temporal evolution of relative variability of documentation time during the project and Figure 3 represents the same data visually. The variability has mostly been downward trending during the pilot project, with July (the most common summer holiday month in Finland), September and December as outliers from the trend. The high general variability points towards heterogeneity in both patient material and workflow between therapists.

Table 3.2: Relative variability of documentation time (%)

Initial	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
IQR/Med	133.0	145.0	113.5	118.4	99.5	119.0	126.7	112.4	85.0	100.5

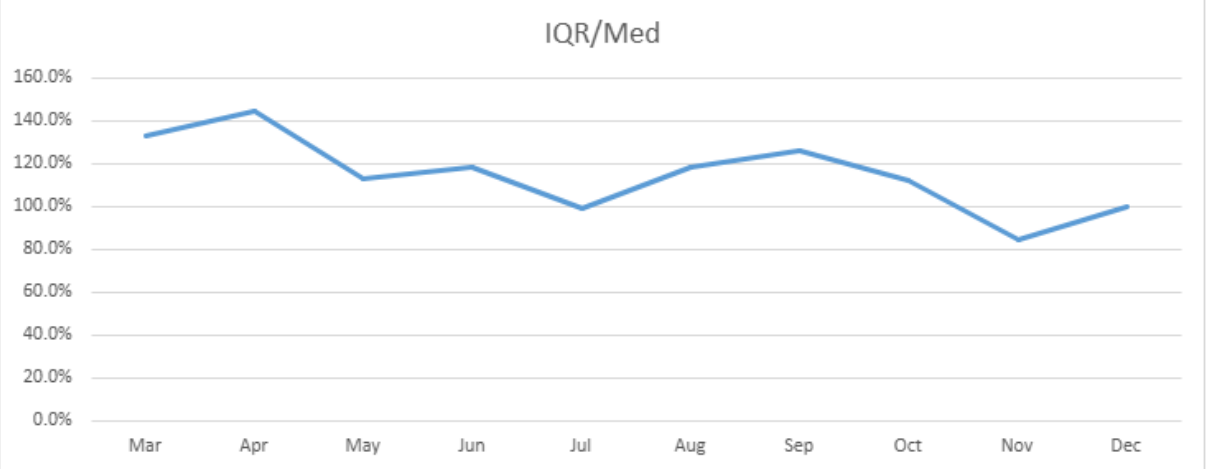


Figure 3.2: Relative variability over time

3.1.2. Temporal Evolution of Documentation Time Distribution

Table 5 shows the distribution of documentation at different percentile ranges. Figure 4 gives a visual representation of the same distribution. The distribution-wide spread follows a very similar trend to the median, peaking in August except for p25 peaking in July. From August to December, the trend has been downward, with the upper-tail (p95) decreasing at a higher rate and the interquartile range narrowing.

Table 3.3: Temporal evolution of documentation time distribution, in seconds

Initial	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
P95	2056	2258	1891	2146	2263	2571	2221	2224	2093	1933
P75	1199	1029	1221	1141	1318	1474	1252	1224	1059	1037
Median	636	560	775	666	790	855	716	746	716	664
P25	352	217	342	353	528	443	345	382	448	370
P5	85	27	34	78	108	119	56	84	129	56

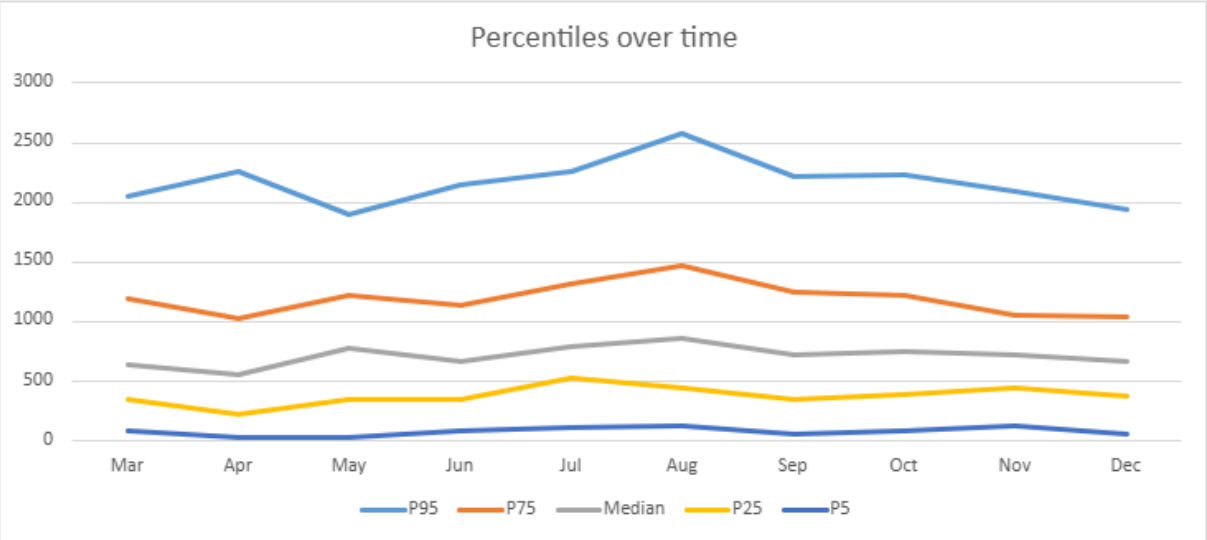


Figure 3.3: Temporal evolution of documentation time distribution, in seconds

3.1.3. Time Stamp Data from the Control Group Survey Period

To avoid temporal confounding of the data, the time stamp data from the same period as the control group data was selected for the comparison. During this period, 33 users produced a total of 340 documentations from initial appointments. The median time per documentation during this period ranged between 11 minutes and 39 seconds and 9 minutes and 48 seconds depending on the filtration. Table 6 shows the key statistics during this period. It is noteworthy that while the reserved time for documentation per initial appointment is 15 minutes (900 seconds), 10 users are showing median times over 1000 seconds in the 30-minute filtration

set. Figures 5, 6 and 7 show the visual representation of both the variability in documentation time for individual users and the variability between users, at different filtration points. In addition to large relative variability (IQR/Median near 1), the data shows very long tails with p95s at 2,6 to 3 times higher than the Median, and p5s at 13 to 14 percent of the median.

Table 3.4: AI scribe group documentation statistics at different filtration points

Gap cutoff	Documen-tations	Mean	Median	Shortest	Longest	IQR	p95	p5	IQR / median	Median of medians
30-minute	340	843.7	698.8	13.1	5093.2	633.2	2104.5	92.9	0.9	678.4
20-minute	340	766	665,7	13,1	2933,3	629,6	1918,2	88,3	0,9	641,7
10-minute	330	667,3	588,2	13,1	2821,3	580,6	1510,6	81,2	1,0	562,3

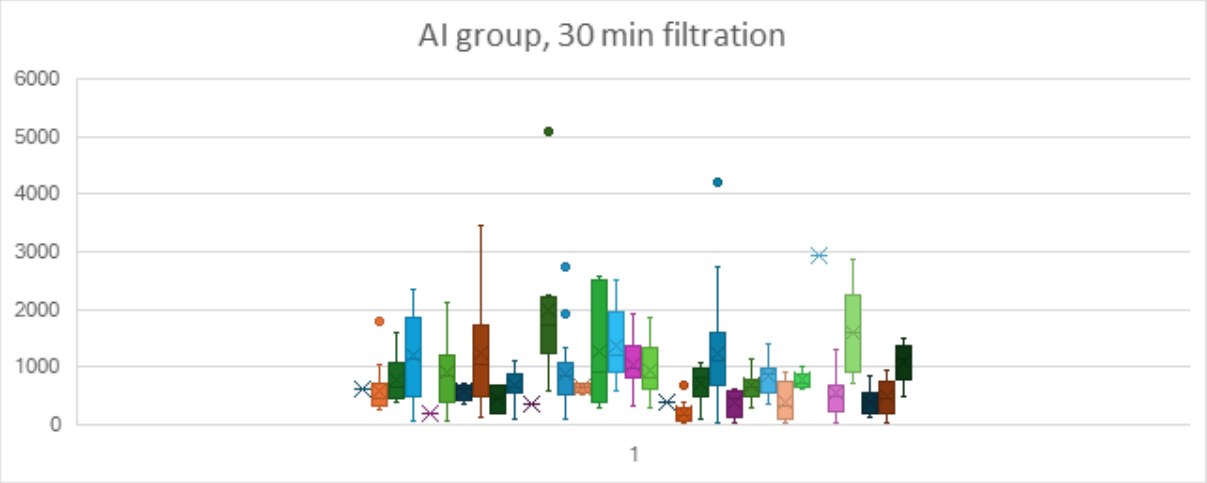


Figure 3.4: Box plot of AI scribe group at 30-minute filtration

Graph showing documentation time in seconds per user

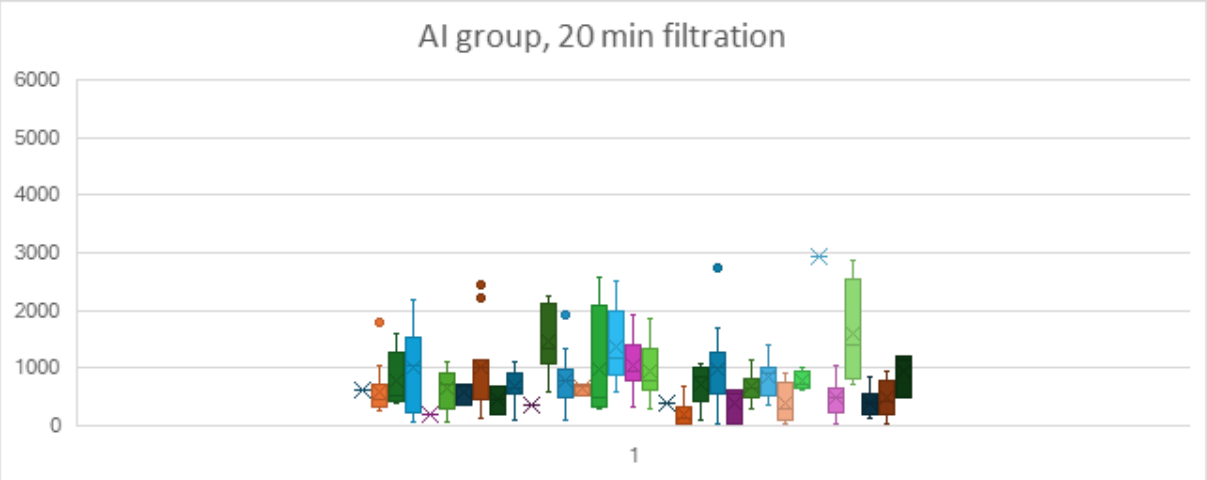


Figure 3.5: Box plot of AI scribe group at 20-minute filtration

Graph showing documentation time in seconds per user

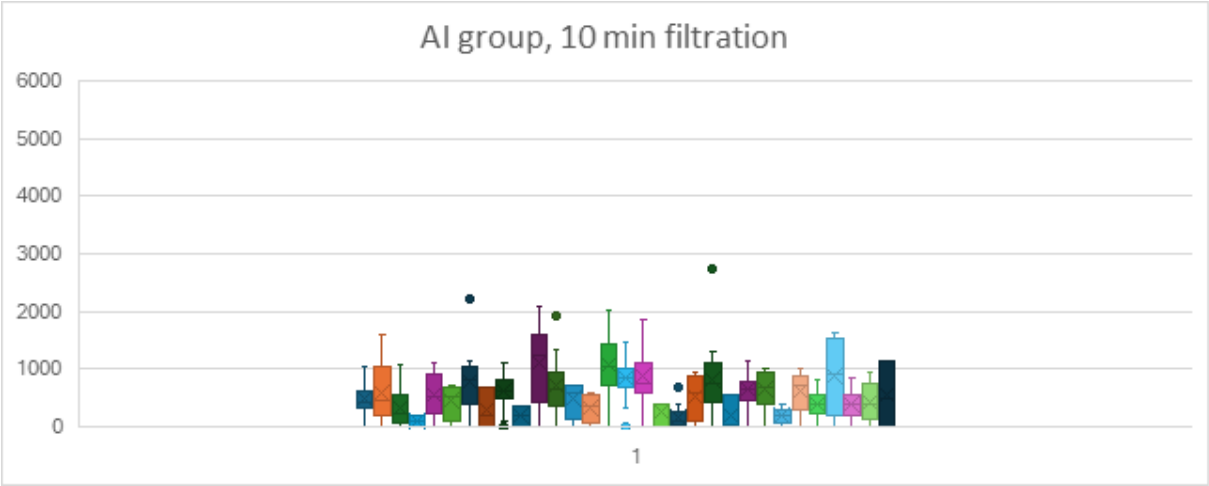


Figure 3.6: Box plot of AI scribe group at 10-minute filtration

Graph showing documentation time in seconds per user

3.2. Control Group Data

The control group data was collected through a survey, asking physiotherapists to log documentation time for initial appointments in minutes, over a two-week period, and then filling out the survey. Table 7 shows the key statistics for the control group.

The control group data shows a much lower variability at 37,5% and shorter tails with p95 at 172,5% and p5 at 50%. Median time spent per text was 16 minutes. Figure 8 shows a visual representation of both the variability in documentation time for individual users and the variability between users.

Table 3.5: Control group documentation statistics

Documentations	Mean	Median	Shortest	Longest	IQR	p95	p5	IQR / median	Median of medians
229	1027	960	300	2400	360	1656	480	0.4	960

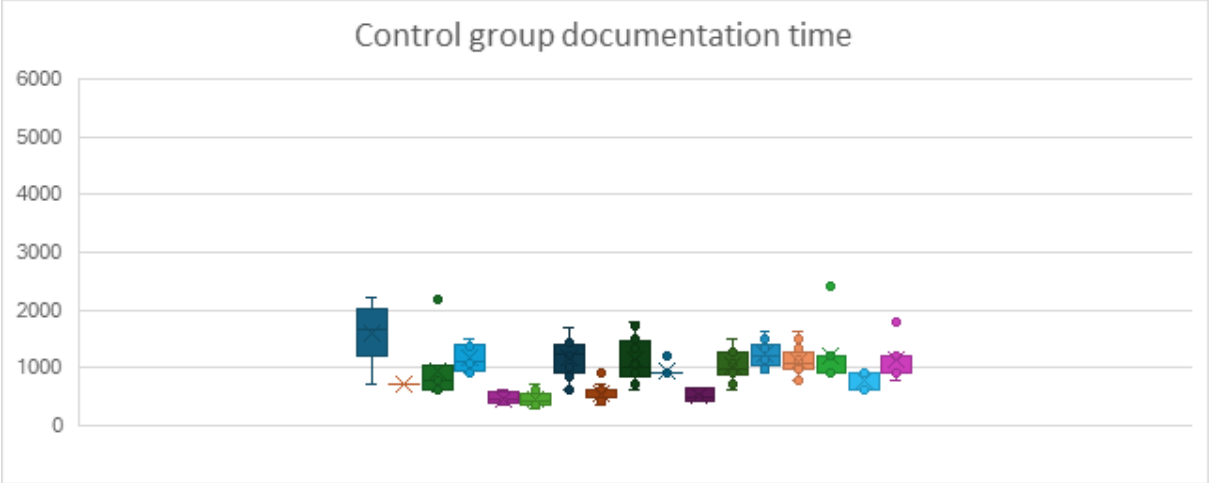


Figure 3.7: Box plot of control group documentation time

Graph showing documentation time in seconds per user

3.3. Time Spent on Documentation, AI Scribe vs Control Group

Time used in seconds per documentation was 960 for the control group. The times at different data filtration points presented in Section 3.6 can be found in Table 8, along with time saved compared to the control group. Time saved ranged from 4,4 minutes to 6,2 minutes, indicating a percentual difference between 27,29 % and 38,75 %.

Table 3.6: Documentation time differentials between control group and AI group at different filtration points

Gap cutoff	Median time (s)	Difference (s)	Difference (min)	Difference (%)
30-minute	698	262	4,4	27,29
20-minute	666	294	4,9	30,63
10-minute	588	372	6,2	38,75

Clinician-level medians were compared using a Mann-Whitney U test. Documentation times were lower in the AI group at all data filtration points, with a detailed breakdown shown in Table 9. Effect size increased at the 10-minute filtration point. The 30-minute filtration did not reach conventional statistical significance, while the 20-minute and 10-minute filtrations did.

Table 3.7: Mann-Whitney U test comparisons between control group and AI group at different filtration points

Gap cutoff	r	Effect size	U	Z	p
30-minute	0,27	Small to moderate	186	-1,94	0,052
20-minute	0,32	Small to moderate	171	-2,233	0,026
10-minute	0,44	Moderate to large	126	-3,06	0,002

Comparing Tables 6 and 7 and Figures 5 to 8 show a considerably larger variability both at clinician level and between clinicians in the AI scribe group compared to the control group. Relative variability and tail behavior remain consistent between different filtrations.

3.4. User Feedback

3.4.1. User Ratings

User ratings were reported on a custom scale from 1 to 5 with the following representations:

1. The documentation required a full rework
2. The documentation required significant modifications
3. The documentation required considerable modifications
4. The documentation required some modifications
5. The documentation required almost no modifications

The average ratings have been trending slowly upwards since August, with the largest increase happening in September. Figure 9 shows a visual representation of the temporal evolution of the average user rating and Figure 10 shows the same representation for the most active users. Comparing the users with over 200 entries each to the overall rating, it can be noted that the trend is very similar even though the average ratings are slightly higher for the most active users.

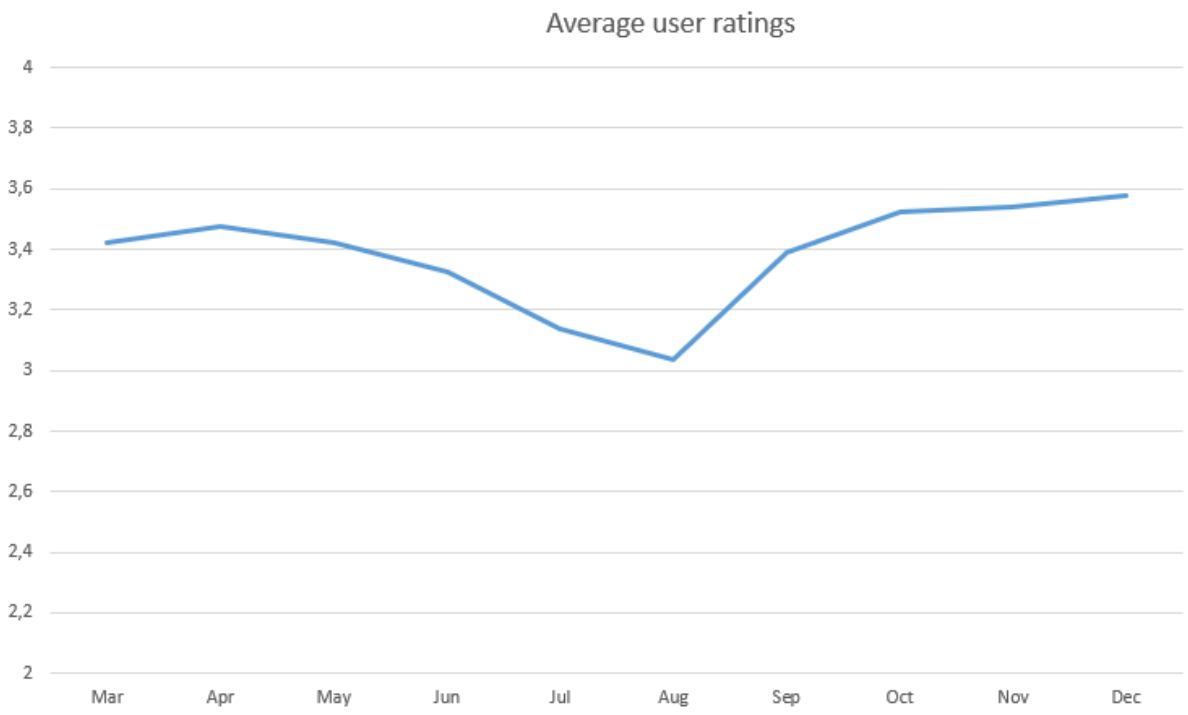


Figure 3.8: Average user rating over time

Rating scale from 1 to 5 described in Section 4.4.1

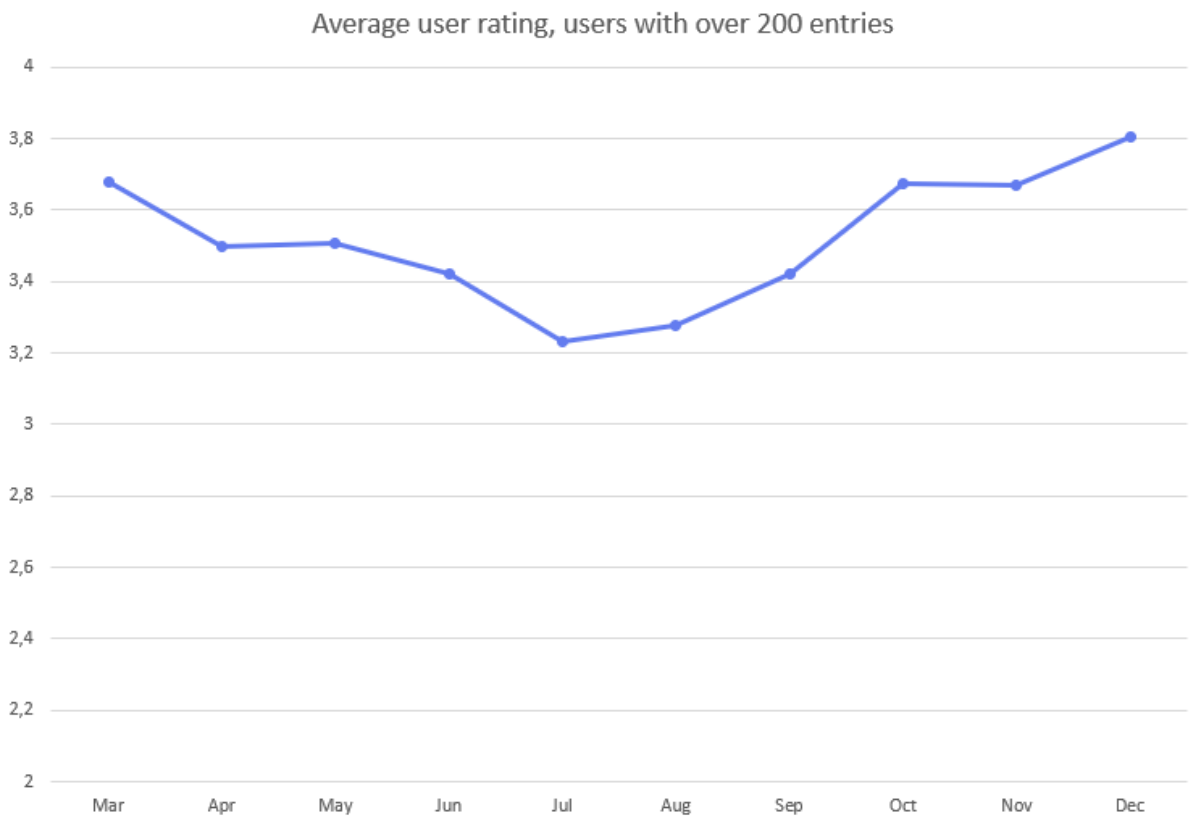


Figure 3.9: Average user rating over time, most active users

Rating scale from 1 to 5 described in Section 4.4.1

3.4.2. Reported Errors

Figure 11 shows the temporal evolution of reported errors by type. The rate of different types of errors occurring in documentations decreased significantly in the early stages, but has levelled off towards autumn, with repetitions proving to be the most persistent problem. It is noteworthy that hallucinations and misinterpretations of information are grouped together in the feedback form, shown under hallucinations in Figure 11.

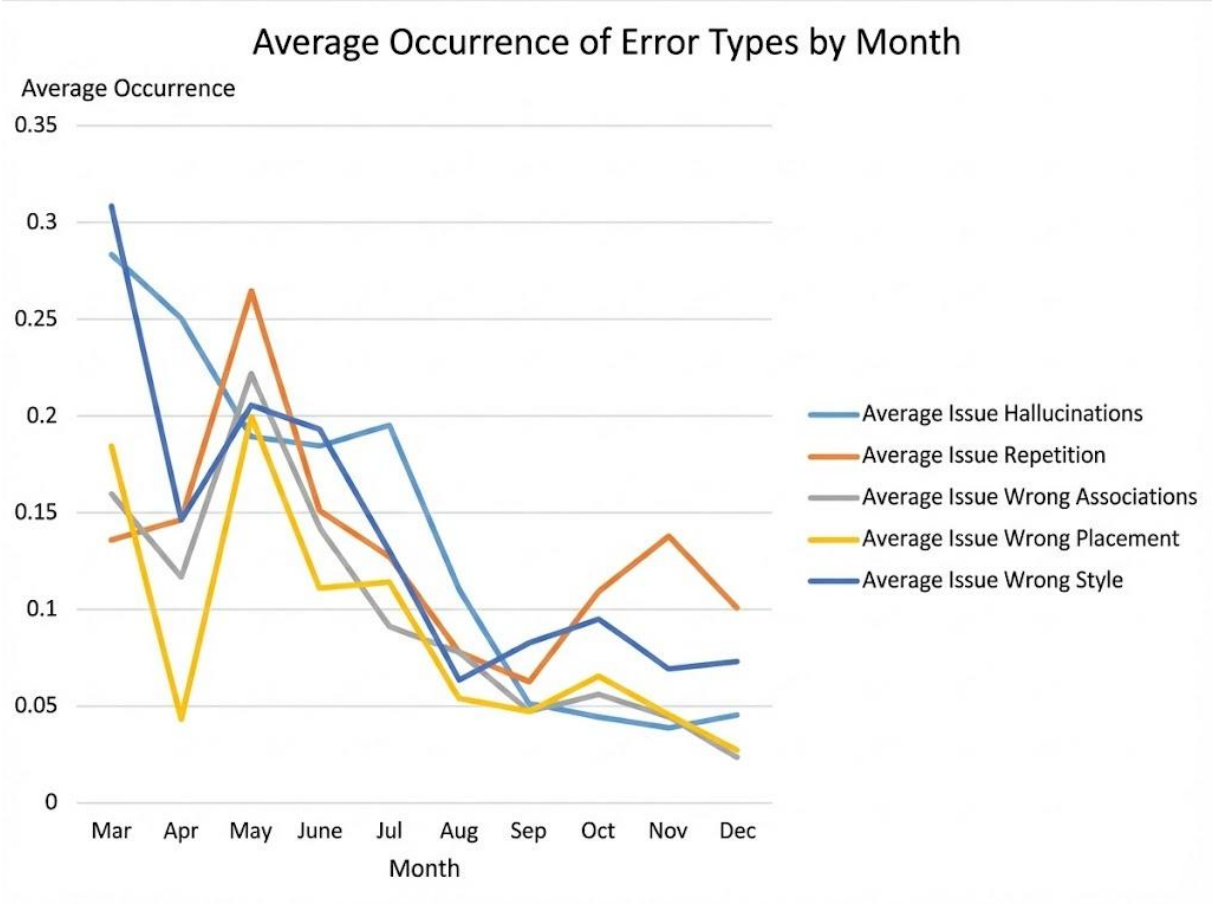


Figure 3.10: Reported errors by type over time

Excel graph modified with Google Nano Banana to remove localization artefacts

3.5. Bi-Weekly User Questionnaire

Despite regular reminders at the Bi-weekly sparring sessions, the response rates remained low for the Bi-Weekly questionnaire. The questionnaire contained 4 questions, asking about the user experience during the past 2 weeks: Figure 12 shows question 1, “How has the AI Scribe affected time used on documentation” (decreased/increased/no effect). Figure 13 shows question 2, “How big that effect has been in minutes per day on average”. Figure 14 shows question 3, “How has the scribe affected therapeutic alliance” (1-5 going from significant negative to significant positive impact). Figure 15 shows question 4, “How the AI scribe has affected job satisfaction” (same 1-5 scale as the previous question).

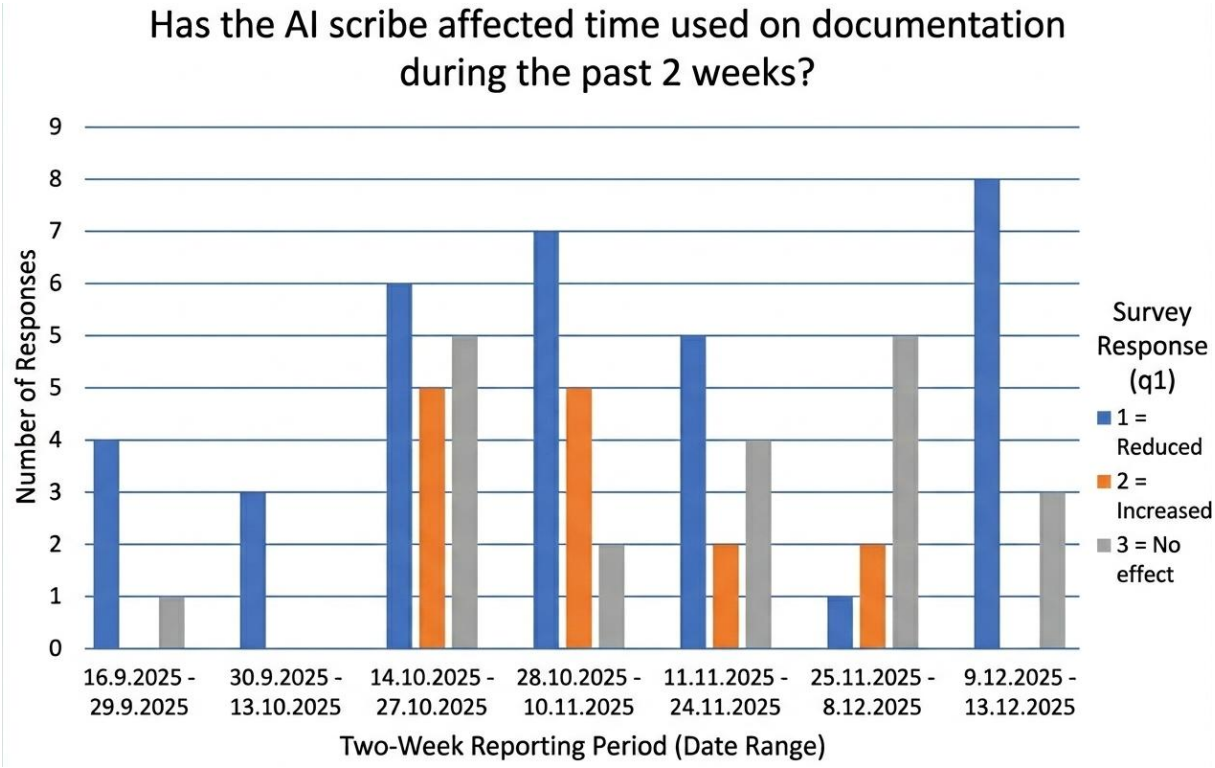


Figure 3.11: Bi-weekly questionnaire Q1

Excel graph modified with Google Nano Banana to remove localization artefacts

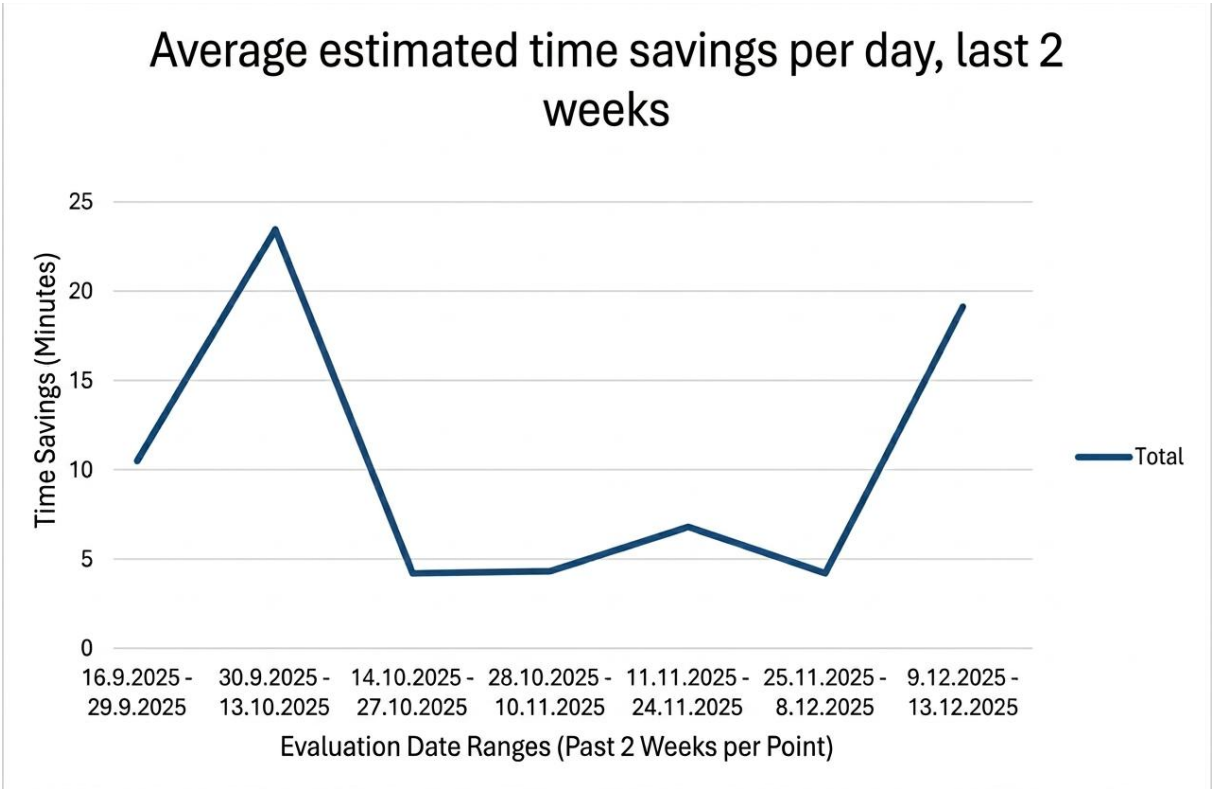


Figure 3.12: Bi-weekly questionnaire Q2

Excel graph modified with Google Nano Banana to remove localization artefacts

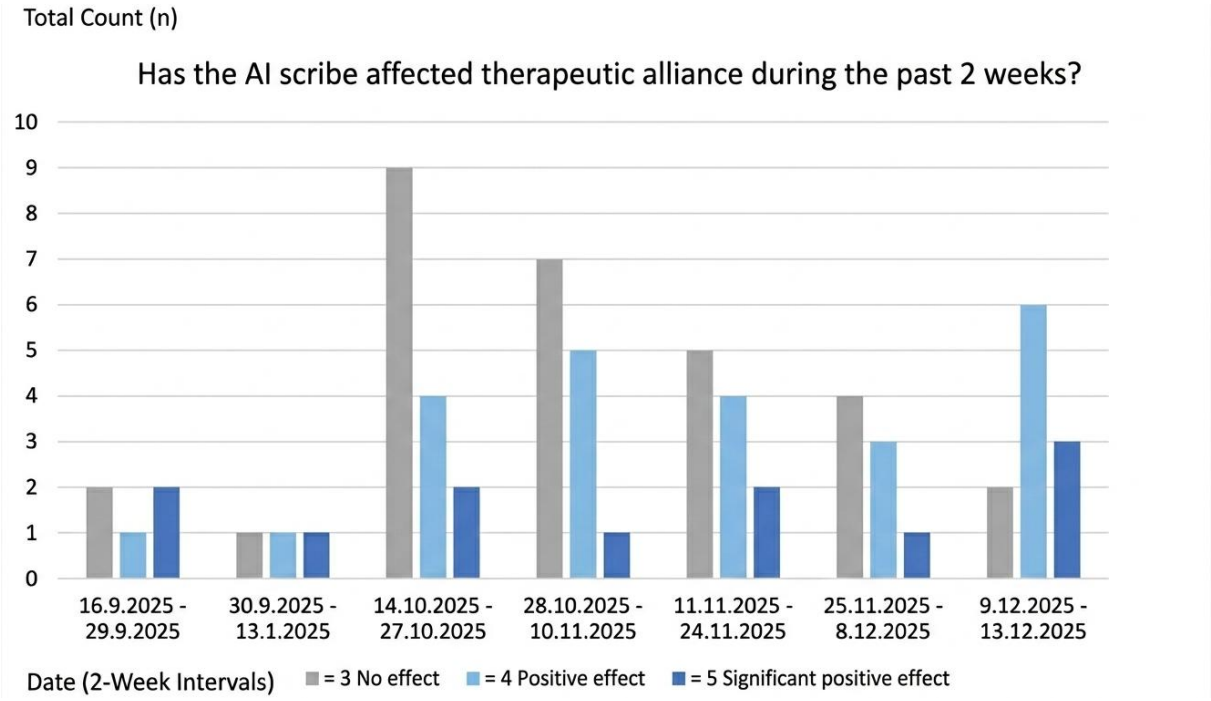


Figure 3.13: Bi-weekly questionnaire Q3

Excel graph modified with Google Nano Banana to remove localization artefacts

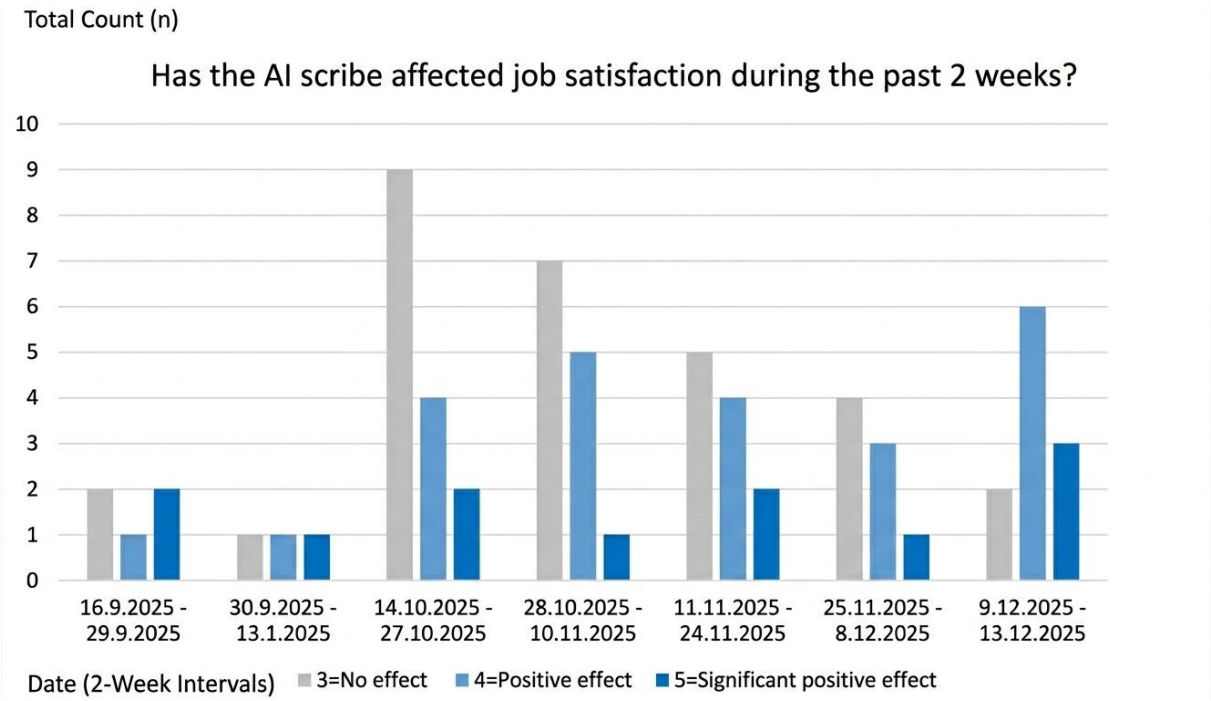


Figure 3.14: Bi-weekly questionnaire Q4

Excel graph modified with Google Nano Banana to remove localization artefacts

While the sample size is small, it is worth noting that nobody reported a negative impact on therapeutic alliance or job satisfaction using the AI scribe. Reported time saved per day varied between 4,16 minutes and 23,33 minutes during the period the questionnaire was up.

4. Discussion

To make sure that the time stamp data doesn't reflect a positive bias, only events with a break over 30 minutes between them were removed in the temporal trend data. As can be seen in Section 4.1, this means that some very long documentation times were included in the dataset, with the longest included time being 102,5 minutes. Similarly, while the very shortest total times (<10 s) were filtered out of the results, it is likely that many of the documentation times close to 10 seconds are not fully edited and copied entries. The effect of this is mitigated by using median values for comparisons.

While spring and summer show shorter documentation times compared to early autumn, the early models could not produce good entries about physical examination. In practice, this meant that the therapists created that part of the early texts from scratch. While everyone was instructed to do all editing work in the text editor of the AI web client, some users may have copied over the entry without the physical examination and finished editing in the EHR. Looking at the trend for time spent, workflow adjustments and the initial learning curve can be seen as the rise in median documentation time in May and decreases in April and June. Late June to early August is the general summer holiday period in Finland, with most employees taking a 3–4-week continuous leave. The increase in July and August is likely to reflect the fact that users had not fully adjusted new workflow before the holidays and had to reacquaint themselves with the process when returning to work. The average user rating curve for spring and summer is quite similar, with the monthly score trending down until August.

Based on sparring sessions and informal discussions with participants, there was a distinct improvement to the quality of the AI scribe when the speech recognition model was updated in September. This is also reflected by the data showing the largest month-to-month reduction in median documentation time and the largest increase in average user rating between August and September. While causal attribution is not possible in this observational design, it is plausible that this change is related to the updated speech

recognition model. This aligns with earlier research on speech recognition. Hodgson & Coeira (2016) demonstrated that transcription accuracy significantly affects the correction burden and documentation safety of documentation. Although this study examined traditional speech recognition systems rather than modern LLM-based solutions, the dependency on transcription quality for results on downstream documentation remains conceptually relevant.

While the monthly average user rating has been slowly increasing each month from September to December, there was a temporary increase in median documentation time in October-November. This suggests that improvements to documentation quality don't necessarily directly correlate with time spent. Another factor here may be that one of the first focuses during the fall sparring sessions was changing usage patterns so that a larger number of users would be able to get usable documentation for physical status.

Figure 4 in Section 4.1.2 showing the distribution-wide improvement suggests a learning and stabilization effect over time. Especially the disproportionate reduction in upper-tail documentation times suggests an improvement to the handling of complex or problematic cases.

4.1. Time Spent on Documentation

Outpatient physiotherapy in the public healthcare system has a very heterogeneous clientele. This can be seen from the large variation in documentation time by single therapists in both the AI scribe- and control groups. Comparing the data shown in Section 4.1.3 and Section 4.2, the AI scribe group presents a significantly larger variability, especially in time used per documentation per therapist, but also between therapists. While this is likely to in part be due to inaccuracies caused by the logging method (it is not always possible to complete the documentation process directly after finishing the physiotherapy session), an assumption supported by the fact that the difference is reduced at tighter filtration points, it is likely that this also highlights that the AI scribe is sometimes able to produce text very close to what is required, while other cases still require a lot of manual editing of the text. The large differences in documentation time between therapists in the AI scribe group also raises the question of workflow differences and ability to present important information in a form that the AI can utilize. While the technical aspects of using

the AI scribe are simple, the required changes to workflow may be challenging. The onboarding and sparring sessions have strived to educate clinicians about how to best utilize the AI scribe. This is however a topic with very little well documented material available. Teachings from the physicians' AI scribe pilot project in LUVN have been used, but there was neither prior experience nor available information about best practices for changes to clinician workflow in a physiotherapy setting.

Based on discussions with participants and personal experience, it seems that more complex cases and clients that meander while presenting their history cause more problems for the AI scribe. As these cases naturally tend to result in longer texts regardless of AI assistance, the effect compounds and could explain the larger internal variability in documentation times for the AI scribe group. Even so, many therapists have expressed the value of the AI scribe for more complex cases, as manually parsing a longer or less structured anamnesis causes a higher cognitive load on the clinician when documented without AI assistance.

The Mann-Whitney U test, presented in Section 4.3, demonstrated a statistically significant difference in documentation times between the groups at two of the three data filtration points. Long breaks between events in the timestamp data can occur in a few different ways. Continuous typing with no breaks over 0,5 seconds counts as one event, which is unlikely to be maintained for 10 minutes or more. Reading without editing text produces longer breaks in the timestamp data. Considering the time reserved for the whole documentation process is 15 minutes, it is unlikely for texts to be so long that the reading/reviewing process would take over 10 minutes without any editing occurring. Thus, the most logical explanation for breaks between events longer than 10 minutes is that the clinician has been busy doing something else. As such it can be argued that the 10-minute cutoff filtration is likely to be closest to actual time spent on documentation in the AI group.

The findings on time usage presented in Section 4.3 are similar to the findings of Evans et al. (2025) who reported approximately 39 % reduction in documentation time compared to baseline for the participants.

Based on discussions with participants and looking at problem cases with documentation in the AI scribe group, the large between-clinician variability in this group is likely to be explained largely by workflow differences and personal documentation preferences.

4.2. Reported Errors in AI Documentation

Regarding reported errors in the texts produced (see Figure 11 in Section 4.4.2), it is worth noting that reporting these errors is an optional step in the feedback process, and as such the absolute numbers are potentially too low. From a safety standpoint hallucinations are the most important errors to eliminate. The large drop in reported hallucinations between August and September highlights the importance of speech recognition and adequate sound quality for the outcome. Repetitions have clearly been the hardest problem to solve. Based on feedback during sparring sessions, a large part of the reported repetition relates to the model struggling with identifying whether factors related to functional status are part of the interview or the physical status and thus often being reported in both.

Hodgson & Coeira (2016), note that clinically significant words may be misrecognized or omitted by the speech recognition systems. This problem can be highlighted in the context of this study, where both a less used language in Finnish and a smaller field of healthcare in physiotherapy present unique and specific vocabulary. Errors may also occur during the interpretation of the transcript by the LLM. Systematic reviews of the ambient era such as Alboksmaty et al. (2025), Bracken et al. (2025) and Sasseville et al. (2025) consistently highlight reliability concerns and the necessity of human oversight. While the data collected during the study does not allow separation of where in the process the errors occur, looking at personal transcripts or the transcripts of other participants, it seems that blatant errors related to hallucination and misinterpretation of information often happen at the speech recognition level. Based on reviewed transcripts, contextual interpretation errors such as repetition and misclassification usually seem to occur during the LLM interpretation phase. Omissions were sometimes due to errors in the transcription or interpretation phase but based on transcription reviews a common issue seems to be workflow related. Transcription reviews would often show that information during the recording had been presented in a form in which the spoken information is insufficient to parse what has happened. Phrases such as “the pain is here”, or “I feel pain during this movement” were common in situations

where users reported issues of misinterpretation of which side of the body the cause of the visit was related to or omissions of results in the physical status section of the documentation. This suggests that ongoing user training to establish good workflow related habits and AI literacy is required to gain the full benefits of an AI scribe. For this study a deeper analysis of these errors was not possible due to data safety and privacy limitations – the transcriptions were not saved, any reviews happened occasionally when participants showed examples of typical problems to the researcher rather than systematically. Further research around documentation quality and safety is warranted to generate deeper insights into the reasons behind different types of errors and how to eliminate them from AI scribe documentation.

A problem with stylistic errors is that it is a very subjective issue. One of the goals with the AI scribe for the organization is to unify documentation. Some therapists may be closer to the target style, or more receptive to the change to their previous style, while others may see this as a larger problem.

4.3. Bi-Weekly Questionnaire

The Bi-weekly questionnaire shows a low to somewhat positive effect on therapeutic alliance and job satisfaction, similar to earlier studies such as Evans et al (2025). The reported estimates on documentation time support the findings of the comparison period discussed in Section 4.3. While the statistical significance of these findings alone is low due to the limited number of responses, the results support the findings of both other research and other collected data of this study.

4.4. Language

The results indicate that the use of an AI scribe is feasible in smaller languages. Support for English and Swedish was added in October. Users did not report any distinct differences in the quality of texts between English and Finnish, while texts in Swedish early on were reported to contain some awkward vocabulary. This is likely because of the AI model primarily being optimized for Finnish, indicating that localization tuning is likely to improve results when implementing AI scribes in a smaller language. As the available data did not differentiate in which language documentations were made this study doesn’t do a deeper analysis on how language affects documentation.

4.5. Limitations

A significant factor affecting results are differences in documentation habits between therapists and physiotherapy units within LUVN. The restructuring of the public healthcare system in Finland was implemented from 1.1.2023, and the process of unifying documentation models within the physiotherapy unit are still ongoing. One of the organizational goals with the AI scribe is to unify documentation. This introduces an interpretative challenge for the study. Many clinicians have reported that they make changes to documentation to suit their personal style, and these personal preferences seem to vary significantly between clinicians. Editing behavior may partially reflect a misalignment with the new organizational standards rather than limitations of the AI generated documentation.

The limited response rates to the bi-weekly and control group questionnaires are likely to reflect the general workload-related burden felt by healthcare professionals, discussed in many previous studies such as Sasseville et al. (2025). Optional feedback activities may be deprioritized due to overall workload, especially when perceived benefits are limited. The voluntary nature of participation was a requirement to align with the ethical standards of LUVN. The optionality also introduces the possibility of response bias.

5. Conclusion

This thesis evaluated the implementation of an ambient AI-assisted clinical documentation system in outpatient physiotherapy at Western Uusimaa Wellbeing Services County using a case study approach.

Regarding documentation time, the results suggest that the use of the AI scribe was associated with a reduction in median documentation time compared to the control group. The direction of the effect was consistent across timestamp filtration assumptions, with the magnitude ranging from 27,3 % to 38,8 %. The distribution of clinician-level medians indicates considerable variability between clinicians. This suggests that time savings are influenced by workflow adaptations and documentation preferences of the users.

Regarding job satisfaction and user experience, the user feedback suggests a generally positive trend over time. No negative effects on therapeutic alliance or job satisfaction were reported. The findings support the results of other studies on the subject, even though the low response rates to the questionnaires limit how strong conclusions can be drawn from the data.

This study contributes novel empirical evidence from a public outpatient physiotherapy context, addressing a clear gap in existing research that has largely focused on physicians. It also demonstrates the viability of AI scribes in smaller languages.

5.1. Findings

This study shows that ambient AI scribes are viable in smaller languages for use in public healthcare and that they reduce the documentation load on clinicians. The results also indicate that users require specific training and workflow optimization to realize the full benefits of the technology. With limited additional training, benefits varied significantly between clinicians in the study.

5.2. Limitations

The study is limited by its observational design, voluntary participation and reliance on self-reported data for some outcomes. Despite these limitations, the findings suggest that ambient AI scribes can meaningfully improve documentation experience in physiotherapy, when supported by the appropriate user training and workflow.

5.3. Future Research

Future research should explore more controlled comparisons between clinical workflows. The results suggest that between clinician differences in workflow affect documentation time significantly. Defining these factors would enable better training for clinicians, enabling the full potential of AI scribes.

Another area of further research is the quality of documentation. Comparing unedited AI documentations to raw transcripts and documentation guidelines could provide deeper insight into which areas of the documentation require further improvement. It would also differentiate errors related to verbally provided information and speech recognition from issues related to the LLM-based processing of the information.

The customer perspective on documentation interpretability is another potential area of future research. National guidelines direct documentation towards a form that is easy for patients to understand, while clinician preferences often lean towards terminology that describes status more precisely. It is likely that AI scribes can steer documentation away from individual preferences, making the information more accessible to patients.

6. Bibliographic References and Sources

Alboksmaty, A., Aldakhil, R., Hayhoe, B. W. J., Ashrafian, H., Darzi, A., & Neves, A. L. (2025). The impact of using AI-powered voice-to-text technology for clinical documentation on quality of care in primary care and outpatient settings: a systematic review. In *eBioMedicine* (Vol. 118). Elsevier B.V. https://doi.org/10.1016/j.ebiom.2025.105861

Blackley, S. v., Huynh, J., Wang, L., Korach, Z., & Zhou, L. (2019). Speech recognition for clinical documentation from 1990 to 2018: A systematic review. In *Journal of the American Medical Informatics Association* (Vol. 26, Issue 4, pp. 324–338). Oxford University Press. https://doi.org/10.1093/jamia/ocy179

Bracken, A., Reilly, C., Feeley, A., Sheehan, E., Merghani, K., & Feeley, I. (2025). Artificial Intelligence (AI) – Powered Documentation Systems in Healthcare: A Systematic Review. In *Journal of Medical Systems* (Vol. 49, Issue 1). Springer. https://doi.org/10.1007/s10916-025-02157-4

Coiera, E., Kocaballi, B., Halamaka, J., & Laranjo, L. (2018). The digital scribe. *Npj Digital Medicine*, *1*(1). https://doi.org/10.1038/s41746-018-0066-9

Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). John Wiley & Sons.

Evans, K., Papinniemi, A., Ploderer, B., Nicholson, V., Hindhaugh, T., Vuvan, V., Cowley, N., Tariq, A., & Thomson, H. (2025). Impact of using an AI scribe on clinical documentation and clinician-patient interactions in allied health private practice: perspectives of clinicians and patients. *Musculoskeletal Science and Practice*, *78*. https://doi.org/10.1016/j.msksp.2025.103333

Falcetta, F. S., de Almeida, F. K., Lemos, J. C. S., Goldim, J. R., & da Costa, C. A. (2023). Automatic documentation of professional health interactions: A systematic review. In *Artificial Intelligence in Medicine* (Vol. 137). Elsevier B.V. https://doi.org/10.1016/j.artmed.2023.102487

Hodgson, T., & Coiera, E. (2016). Risks and benefits of speech recognition for clinical documentation: A systematic review. In *Journal of the American Medical Informatics Association* (Vol. 23, Issue e1, pp. e69–e179). Oxford University Press. https://doi.org/10.1093/jamia/ocv152

Leung, T. I., Coristine, A. J., & Benis, A. (2025). AI Scribes in Health Care: Balancing Transformative Potential With Responsible Integration. In *JMIR Medical Informatics* (Vol. 13). JMIR Publications Inc. https://doi.org/10.2196/80898

Lukac, P. J., Turner, W., Vangala, S., Chin, A. T., Khalili, J., Shih, Y.-C. T., Sarkisian, C., Cheng, E. M., & Mafi, J. N. (2025). *A Randomized-Clinical Trial of Two Ambient Artificial Intelligence Scribes: Measuring Documentation Efficiency and Physician Burnout*. https://doi.org/10.1101/2025.07.10.25331333

Ng, J. J. W., Wang, E., Zhou, X., Zhou, K. X., Goh, C. X. le, Sim, G. Z. N., Tan, H. K., Goh, S. S. N., & Ng, Q. X. (2025). Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: a systematic review. In *BMC Medical Informatics and Decision Making* (Vol. 25, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s12911-025-03061-0

Omon, K., Sasaki, T., Koshiro, R., Fuchigami, T., & Hamashima, M. (2025). Effects of Introducing Generative AI in Rehabilitation Clinical Documentation. *Cureus*. https://doi.org/10.7759/cureus.81313

Perkins, S. W., Muste, J. C., Alam, T., & Singh, R. P. (2024). *Improving Clinical Documentation with Artificial Intelligence: A Systematic Review*.

Sasseville, M., Yousefi, F., Ouellet, S., Naye, F., Stefan, T., Carnovale, V., Bergeron, F., Ling, L., Gheorghiu, B., Hagens, S., Gareau-Lajoie, S., & LeBlanc, A. (2025). The Impact of AI Scribes on Streamlining Clinical Documentation: A Systematic Review. In *Healthcare (Switzerland)* (Vol. 13, Issue 12). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/healthcare13121447

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). *Compute Trends Across Three Eras of Machine Learning*. https://doi.org/10.1109/IJCNN55064.2022.9891914

Sosiaali- ja terveysministeriön asetus potilasasiakirjoista 94/2022. (2022). *Suomen säädöskokoelma*. Retrieved from https://www.finlex.fi/fi/lainsaadanto/saadoskokoelma/2022/94

Terveyden ja hyvinvoinnin laitos. (2025). *Potilastiedon kirjaamisen yleisopas* (Version 7.0) https://yhteistyotilat.fi/wiki08/spaces/JULPOKY/overview

Turpin, P. G. (2005). *Transitioning From Paper to Computerized Documentation Computerized Systems in Nursing*. https://doi.org/https://doi.org/10.1097/00001610-200501000-00016

van Buchem, M. M., Boosman, H., Bauer, M. P., Kant, I. M. J., Cammel, S. A., & Steyerberg, E. W. (2021). The digital scribe in clinical practice: a scoping review and research agenda. In *npj Digital Medicine* (Vol. 4, Issue 1). Nature Research. https://doi.org/10.1038/s41746-021-00432-5

Yin, R. K. (2018)**.** *Case study research and applications: Design and methods* (6th ed.). SAGE Publications.

7. Annexes

7.1. Annex A – Literature Search Strategy

A structured literature search was conducted to identify empirical research on ambient AI, digital scribes and speech recognition technologies used for clinical documentation, focusing on rehabilitation and physiotherapy settings.

Google Scholar and PubMed were selected as the databases for the literature search. The final search was performed on December 1, 2025. No lower date restriction was applied to the search.

7.1.1. Search Strings

Two searches were performed in Google Scholar, with the following search strings:

Search 1: ("ambient AI" OR "ambient intelligence" OR "automatic speech recognition" OR "voice-to-text" OR "speech recognition")

AND ("clinical documentation" OR "electronic health records" OR "clinical notes")

Search 2: ("clinical documentation" AND physiotherapy OR "physical therapy" OR rehabilitation OR musculoskeletal)

AND ("speech recognition" OR "ambient AI" OR "natural language processing")

One search was performed in PubMed, with the following search string:

(("Speech Recognition Software"[Mesh] OR "speech recognition"
OR "automatic speech recognition" OR "voice-to-text"
OR "dictation software" OR "ambient intelligence"
OR "ambient AI" OR "natural language processing"
OR "large language model" OR "generative AI"))
AND
(("Electronic Health Records"[Mesh] OR "electronic health record"
OR "EHR" OR "clinical documentation"
OR "clinical notes" OR "medical documentation"))
AND
(("Physical Therapy Modalities"[Mesh] OR physiotherapy
OR "physical therapy" OR musculoskeletal

OR rehabilitation OR "rehabilitation medicine"))

AND

(feasibility OR usability OR "workflow efficiency"

OR "user acceptance" OR "human factors"

OR "socio-technical" OR implementation)

Medical Subject Headings (MeSH) were used where applicable in PubMed to improve search sensitivity and specificity.

7.1.2. Screening and Selection Process

Search results were screened in three stages:

1. Title screening for relevance to AI-assisted clinical documentation in healthcare.
2. Abstract screening to determine if the study focused on a clinical setting and to determine what type of technology was evaluated.
3. Full-text review for final inclusion.

Additionally, reference lists of relevant studies were screened to identify additional sources. AI-based tools (ChatGPT 5.1 and Google Gemini 3.0) were used to assist in identifying any research gaps in the initial searches. All included articles were verified through database searches and full text reviews to ensure relevance and eligibility.

7.1.3. Inclusion Criteria

To be included, a study had to report empirical findings on speech recognition, digital scribes, ambient AI or LLM-based documentation tools. The study had to assess real or simulated clinical settings. If assessing a general healthcare setting, only systematic reviews or RCTs were included. For rehabilitation-specific studies, all studies were screened. Additionally, the study needed to assess outcomes related to documentation time, workflow efficiency, cognitive load or burnout, documentation quality, usability or implementation.

7.1.4. Exclusion Criteria

Studies were excluded if they focused solely on the technical aspects without clinical evaluation, did not relate to healthcare documentation or were not available in English.

7.1.5. Results of the Search

The searches yielded a body of literature covering early speech recognition systems and more recent LLM-based ambient AI documentation tools. After screening and eligibility assessment, five systematic reviews, two empirical studies from a rehabilitation setting and one large RCT were included in the literature review.

7.2. Annex B – Research Plan

7.2.1. Working title

Evaluation of Ambient AI-Based Clinical Documentation in Physiotherapy: A case study of Western Uusimaa Wellbeing Services County

7.2.2. Relevance

Physiotherapy specific research on the use of ambient AI-assisted documentation is very limited, and research on ambient AI solutions using the Finnish language are even more so. Additionally, this study aims to evaluate the effects of a novel solution tailored to the needs of Western Uusimaa Wellbeing Services County (LUVN).

The Ambient AI scribe pilot at LUVN started in the spring of 2024 for physicians and was expanded to physiotherapy in March 2025. The pilot provides a unique opportunity to evaluate both how well the technology works in a real-world clinical setting, and how different improvements to the model and workflow of professionals affect the output.

7.2.3. Problem Definition and Research Gap

Existing studies on AI-assisted documentation have primarily focused on physicians. While outpatient physical therapy shares some of the structure for documentation, there is usually a heavier focus on movement and function in the form of observations and clinical tests. This may challenge voice-to-text models as more of the relevant information might only be implied, not directly voiced out loud.

The primary focus of this thesis is to improve understanding of how ambient AI impacts the documentation burden and workflow of physiotherapists in an outpatient setting.

7.2.4. Aim and objectives

The aim of this thesis is to evaluate the implementation of ambient AI-assisted clinical documentation in physiotherapy services at LUVN using a Case Study Research approach.

Objectives:

1. To assess how use of ambient AI affects time spent on documentation
2. To evaluate how use of ambient AI affects job satisfaction
3. To evaluate how use of ambient AI affects patient interaction
4. To generate insight into what kind of technical and workflow-related improvements lead to better quality of documentation

7.2.5. Data collection and Analysis

Data sources:

- Bi-Weekly, anonymous questionnaire data collected by the company developing the ambient AI model, with the following questions:
    - Estimated time spent on documentation with ambient AI-assistance, compared to without and an estimation of minutes saved/wasted per day.
    - Perceived effect on job satisfaction while using ambient AI
    - Perceived effect on patient interaction while using ambient AI
- Weekly statistics collected from anonymous feedback of each generated documentation, collected by the ambient AI company and the RDI unit at LUVN, with the following information:
    - Quality of the text, reported by how much editing was required (1= required a lot of editing, 5= required almost no editing)
    - Amount of users
    - Amount of generated documentation

- Pseudonymized time stamp data to measure time spent editing AI generated documentation by the professionals
- Anonymously collected, self-reported time spent per documentation of a comparison group of physiotherapists at LUVN
- Dates of larger improvements to the ambient AI model and what type of improvement was made, reported by the company developing the model
- Dates of additional training and sparring sessions provided to the users, reported by the author

Quantitative data will be analyzed to evaluate how documentation efficiency and quality evolve over time. The data will be compared across time periods corresponding to significant improvements to the ambient AI model and training interventions. This comparison will help identify relationships between system refinements, user competence development and outcomes. Descriptive and inferential statistical methods will be used to assess and visualize these changes.

Data triangulation will be used to strengthen the validity of the results on potential time saved using ambient AI-assisted documentation.

7.2.6. Scope and Limitations

The study focuses on the use of a specific implementation of ambient AI within Western Uusimaa Wellbeing Services County, and use is limited to clinical visits of outpatient physiotherapy, focusing on musculoskeletal issues. The results may not be generally applicable to other user groups, settings, or AI solutions.

Due to the scope of the thesis, the collected data will be from a relatively short period and thus may not show the full picture of the development process.

As the author of this thesis is directly involved in both managing the project and using the ambient AI model, there is a risk of bias in any findings not directly based on quantitative data.

7.2.7. Ethical considerations

Research approval will be applied for according to LUVN's internal research authorization process.

No direct or indirect patient data will be collected for this thesis. Data already being collected by other parties will be provided to the author either in anonymous or pseudonymized form.

Comparison group data will be collected anonymously, and participation will be voluntary and based on informed consent.

Data will be stored securely according to GDPR and LUVN's data protection policies.

Expected outcomes

- Empirical evidence on the benefits and limitations of ambient AI-assisted documentation in physiotherapy.
- Practical insights for healthcare management on AI adoption.
- Insights into the design process of ambient AI-assisted documentation systems

Table 7.1 Timeline

Phase	**Timeframe**	**Activities**
Literature review and research design	June- October 2025	Review theoretical background, finalize study design
Data collection preparation	October – November 2025	Research approval, comparison group recruitment, tool setup
Data collection	November – December 2025	Conduct comparison group questionnaire, collect data from partners
Analysis and synthesis	December 2025	Data coding and analysis
Writing and submission	December 2025- January 2026	Draft, revise and finalise thesis

7.2.8. Research group

Researcher: John Sundberg, [Contact information redacted]

Supervisor: Jyri Rajamäki, Yliopettaja, Laurean Ammattikorkeakoulu, [Contact information redacted]

LUVN contact person: Sanna Vuorinen, Palveluyksikköpäällikkö, Fysioterapiapalvelut, [Contact information redacted]

7.2.9. Key references

Duggan, M. J., Gervase, J., Schoenbaum, A., Hanson, W., Howell, J. T. III, Sheinberg, M., & Johnson, K. B. (2025). Clinician experiences with ambient scribe technology to assist with documentation burden and efficiency. JAMA Network Open, 8(2), e2460637. https://doi.org/10.1001/jamanetworkopen.2024.60637

Ng, J. J. W., Wang, E., Zhou, X., Zhou, K. X., Goh, C. X. L., Sim, G. Z. N., Tan, H. K., Goh, S. S. N., & Ng, Q. X. (2025). Evaluating the performance of artificial intelligence–based speech recognition for clinical documentation: A systematic review. BMC Medical Informatics and Decision Making, 25, Article 236. https://doi.org/10.1186/s12911-025-03061-0

Yin, R. K. (2018). *Case study research and applications: Design and methods* (Sixth edition.). SAGE.

7.3. Annex C – Informed Consent Form

Tutkimusluvan diaarinumero: 4363/13.01.00/2025

Tutkimuksen nimi:
Evaluation of Ambient AI-Based Clinical Documentation in Physiotherapy: A Case Study of Western Uusimaa Wellbeing Services County

1. Tutkimuksen tausta ja tarkoitus
Tämän tutkimuksen tavoitteena on arvioida tekoälyavusteisen kirjaamisen käyttöä fysioterapiassa ja sen vaikutuksia dokumentoinnin tehokkuuteen ja laatuun. Tutkimus

toteutetaan osana Managing Digital Transformation in the Health Sector (ManagiDiTH) -maisteriohjelman pro gradu -tutkielmaa. Tässä osatutkimuksessa kerätään vertailuaineistoa fysioterapeuteilta, jotka eivät käytä tekoälyavusteista kirjaamisratkaisua. Kerättävä tieto koostuu itsearvioiduista kirjaamiseen käytetyistä ajoista kahden viikon ajalta.

2. Osallistumisen vapaaehtoisuus

Tutkimukseen osallistuminen on täysin vapaaehtoista. Voit kieltäytyä osallistumasta tai keskeyttää osallistumisesi milloin tahansa ennen vastausten lähettämistä ilman mitään seurauksia.

3. Kerättävät tiedot ja aineiston käsittely

Kyselyssä kerätään ainoastaan anonyymiä tietoa Microsoft Forms - lomakkeen kautta. Kysely ei sisällä henkilötietoja eikä yksittäisiä vastauksia voida yhdistää tiettyyn henkilöön. Vastaukset tallennetaan Länsi-Uudenmaan hyvinvointialueen (LUVN) suojattuun pilvipalveluun, ja tietoja käytetään ainoastaan tämän tutkimuksen analysointiin ja raportointiin. Tietoja ei luovuteta ulkopuolisille.

4. Tietosuoja ja aineiston säilytys

Kaikki aineisto käsitellään tietosuoja-asetusten (GDPR) ja tutkimuseettisten periaatteiden mukaisesti.

- Tietoja säilytetään LUVN:n suojatussa pilvipalvelussa.
- Aineisto poistetaan kokonaisuudessaan tutkimuksen valmistuttua, viimeistään vuoden 2026 loppuun mennessä.
- Koska tiedot ovat täysin anonyymejä, yksittäisiä vastauksia ei voida myöhemmin poistaa tunnistettavasti.

5. Riskit ja hyödyt

Tutkimukseen osallistumisesta ei aiheudu osallistujille riskejä tai haittaa. Tutkimuksen tuloksia hyödynnetään tekoälyavusteisen kirjauksen työkalun arvioinnissa ja kehittämisessä.

6. Tulosten raportointi

Tulokset raportoidaan ryhmätasolla siten, ettei yksittäisiä vastaajia voida tunnistaa. Tutkimuksen tuloksia voidaan hyödyntää akateemisessa raportoinnissa ja esityksissä ilman yksilöiviä tietoja.

7. Lisätietoja tutkimuksesta

Jos sinulla on kysyttävää tutkimuksesta, voit olla yhteydessä:

Tutkija: John Sundberg, [Contact information redacted]

LUVN yhteyshenkilö: Sanna Vuorinen, [Contact information redacted]