Evaluating Ambient AI for Clinical Documentation: A Mixed-Methods Study

Moss Mille [a,*], Bagger Stine [a,*], Riise Louise [a,*], Schultz Robert [a,*],

[a] *Department of Health Science and Technology, Aalborg University, Gistrup, Denmark.*
* These authors contributed equally to the work

ARTICLE INFO

Keywords:
Ambient AI
Large Language Model
Speech-to-text
Conversationsal AI
AI scribe
Documentation

ABSTRACT

Background & Objective: Documentation in electronic health records increases clinician workload and burnout. Ambient AI scribes may reduce this burden by generating draft notes, but challenges remain. This study investigates how ambient AI scribes affect documentation work in clinical practice.
Methods: A mixed-methods explanatory sequential design was used. Two clinicians, a psychologist and a nurse, participated. Quantitative data included documentation time and error rates, analysed using Mann-Whitney U tests and Pearson's and Spearman's correlations. Qualitative data included a semistructured interview, analysed thematically following Braun and Clarke.
Results: Quantitative analyses showed increased documentation time for the nurse (7.5 vs. 21.5 min, $p = 0.024$), while no difference was found for the psychologist (62.0 vs. 86.4 min, $p = 0.190$). A learning effect was observed only for the psychologist ($\rho = –0.689$, $p = 0.040$, 95 % CI [–0.932, –0.022]). Mean error rates were higher for the psychologist (34.11 (SD 12.33)) than the nurse (9.33 (SD 3.50). Error rate was not associated with documentation time or number of sessions.
Qualitative analyses indicated that ambient AI was intuitive and helpful for structuring draft notes, however frequent errors, missing information, and hallucinations limited trust and prevented time savings.
Conclusion: Ambient AI did not improve documentation efficiency, was associated with ambivalent user satisfaction, and did not demonstrate sufficient effectiveness. Further research is necessary due to study limitations.

1. Introduction

Documentation in electronic health records (EHRs) constitutes a central factor contributing to clinician burnout, stress, and overall workload [1–13]. This documentation burden is primarily driven by time consuming documentation requirements, complex workflows, and limited system usability [2,8,11–17]. To address these challenges, several studies have explored ambient artificial intelligence (AI) scribes as a potential solution to reduce documentation burden [5,12,18–27].

Ambient AI scribes are AI-driven speech-to-text systems, often leveraging generative AI based on large language models (LLMs), that automatically capture, transcribe, and structure clinical consultations into draft notes for review and editing by clinicians (27).

Several studies have found that ambient AI can significantly reduce the time clinicians spend on documentation [18,21,22,24,26,27], and some studies also report additional benefits, including reduced workload, lower burnout, and improved job satisfaction [18,23–26].

While ambient AI is promising in alleviating the documentation burden, challenges remain. These include variability in individual outcomes, the need for post-editing of AI-generated draft notes, potential language barriers, and required adjustments to workflows and staff training to ensure effective implementation [18,19,21,25].

Given these potential benefits and remaining challenges, this study aims to investigate how ambient AI affects documentation work in clinical practice.

2. Methods

2.1 Study design

A mixed-methods design using an explanatory sequential approach was applied to investigate how ambient AI affects documentation work in clinical practice. The quantitative phase assessed documentation errors and time usage during clinical consultations, followed by a qualitative phase that explored user satisfaction through an individual interview (Fig. 1). Data were collected from November 14th to December 10th.

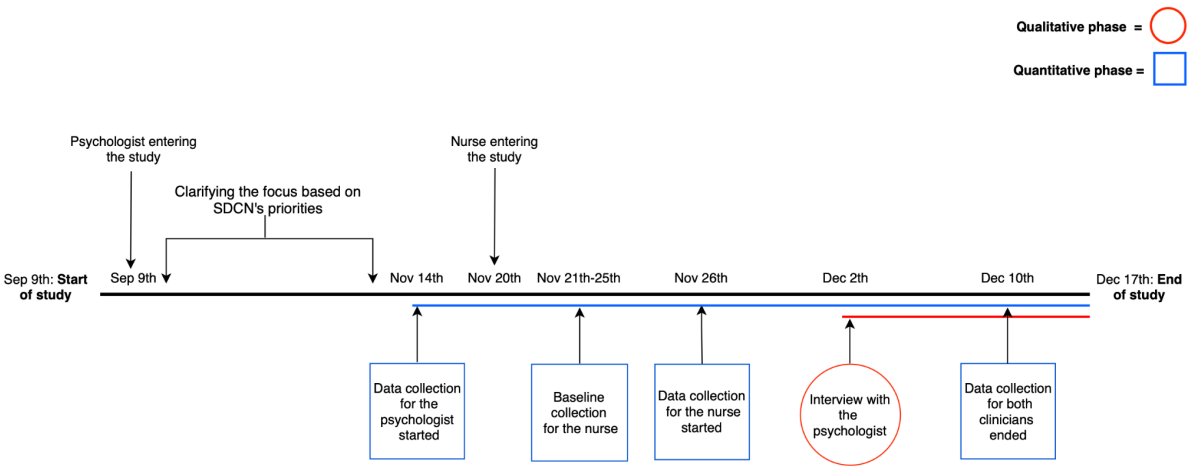


Fig. 1: *Study timeline showing key milestones and the periods for the qualitative (red) and quantitative (blue) phases, including clinician entry, data collection, and the interview.*

2.2 Setting

The study was conducted at the Steno Diabetes Center North Denmark (SDCN), located at Aalborg University Hospital [28]. SDCN piloted the ambient AI scribe Corti in semi-structured patient consultations [29].
Corti is an ambient AI documentation tool that generates structured draft notes based on the dialogue between clinician and patient.

2.3 Participants

Two clinicians participated in the study: one psychologist (from September 9th) and one nurse (from November 20th). Corti was piloted during the psychologist's in-person consultations with children and adolescents with type 1 diabetes, as well as in separate consultations with their parents. For the nurse, Corti was piloted during telephone consultations with socially vulnerable adults with poorly controlled diabetes. All consultations were semi-structured. Quantitative data were collected from both clinicians, while qualitative data were collected only from the psychologist.

2.4 Theoretical Framework

The evaluation was grounded in the concept of usability, as defined by ISO 9241–11. Usability was assessed across three core dimensions [30].
Efficiency: The relationship between performance and the resources required to achieve it.
Effectiveness: The extent to which users are able to achieve their goals accurately and completely using the system.
Satisfaction: The user's subjective experience of the system's usability and work-related benefits.

2.5 Assessment of Efficiency and Effectiveness

Quantitative data were collected to evaluate *efficiency* and *effectiveness*.

Efficiency was assessed by documentation time per clinical note, both with and without using Corti, including potential learning effect. For the psychologist, baseline estimates without Corti were based on two thirds of consultation time typically spent on documentation, while for the nurse, documentation time without Corti was measured during November 21st to November 25th. Documentation time with Corti was measured by both clinicians and noted in the error classification scheme (Appendix A).

Effectiveness was assessed by identifying errors in Corti-generated draft notes. Two initial note reviews were conducted in which the psychologist edited the notes while the authors, acting as non-participatory observers, documented all errors identified by the psychologist. Two authors attended the first session, and two different authors attended the second. These errors were categorised to develop an internal error classification scheme, which was refined in collaboration with the psychologist. Each identified mark was counted as one error. The scheme (Appendix A) was piloted in two documentation sessions with only minor revision therefore, pilot data were included in the analysis.

2.6 Quantitative Analysis

Normality was assessed using the Shapiro-Wilk test.

Efficiency was assessed using Mann-Whitney U tests for documentation time comparisons. Learning effects were analysed using Spearman's rho and Pearson's correlation.

Effectiveness was assessed using descriptive statistics of error rates. Associations

between error rate, documentation time, and session number were analysed using Spearman's rho and Pearson's correlation.

A significance level of $p < 0.05$ was applied. All statistical analyses were conducted using SPSS (version 31.0.0.0(117)), and all figures were generated in MATLAB (version 24.2.0 (R2024b)).

2.7 Assessment of Satisfaction

Satisfaction was assessed using qualitative data. A semi-structured interview was conducted with the psychologist to explore the experience using Corti in clinical documentation. The interview guide was developed deductively with a focus on *satisfaction*, while also including questions on *efficiency, effectiveness*, and workflow. The interview questions were sent to the psychologist 72 hours in advance. Questions related to *efficiency* and *effectiveness* were based on preliminary quantitative results and addressed frequent error types, changes in error rate over time, and time spent on documentation. A pilot interview was conducted to refine the guide prior to the interview (Appendix B).

The interview lasted 92 minutes and took place in the psychologist's clinical workspace. At the beginning of the interview, the psychologist was briefed on its purpose, structure and signed informed consent. Two authors participated: one conducted the interview, and the other managed timekeeping and ensured coverage of key questions. The interview ended with a debriefing, allowing the psychologist to add or clarify points. The interview was audio-recorded and transcribed verbatim by the same authors according to predefined transcription guidelines (Appendix C).

2.8 Qualitative Analysis

Interview data were analysed using thematic analysis following Braun and Clarke's framework [31]. An inductive approach was adopted, allowing themes to emerge from the data without reliance on predefined coding categories [31]. Coding and theme development were carried out by the two authors who have not participated in the interview. The analysis proceeded through: (1) familiarisation with the data; (2) systematic generation of initial codes; (3) searching for themes by organising related codes into preliminary thematic clusters; (4) the preliminary themes were reviewed to ensure clarity. As the analysis was based on a single interview, the focus was on interpretation; (5) defining and naming themes to capture the central patterns in the psychologist's experiences, including a credibility check with the psychologist; and (6) reporting the final themes in the Results section, supported by quotations [31].

3. Results

3.1 Quantitative Results

A total of 16 error classification schemes were collected: 6 from the nurse and 10 from the psychologist. 1 scheme from the psychologist was excluded due to Corti not being activated during the entire consultation, resulting in data from 15 schemes being included in the analysis. All results are presented in Table 1 and 2.

3.1.1. Efficiency

Documentation time with vs. without Corti

For the psychologist, mean documentation time was 62.0 minutes (SD = 14.83; range 40–80) at baseline and 86.4 minutes (SD = 38.82; range 45–175) with Corti ($p = 0.190$) (Fig. 2).

For the nurse, mean documentation time was 7.5 minutes (SD = 0 at baseline and

21.5 minutes (SD = 3.83; range 16–26 with Corti (p = 0.024) (Fig. 2).

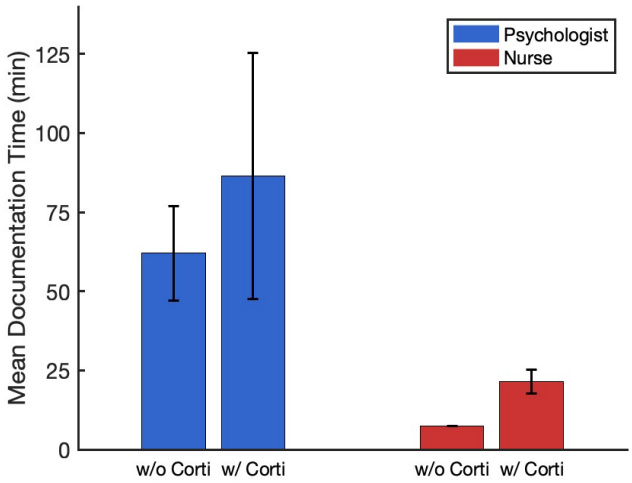


Fig. 2: *Mean documentation time with and without Corti for the psychologist (blue) and the nurse (red). Bars show mean values, and error bars indicate standard deviations (SD).*

Learning effect

For the psychologist, a significant decrease in documentation time across sessions was observed (Spearman's ρ = –0.689, p = 0.040, 95 % CI [–0.932, –0.022]).

For the nurse, no significant association between documentation time and session number was observed (Pearson's r = –0.655, p = 0.158, 95 % CI [–0.958, 0.334]) (Fig. 3).

3.1.2 Effectiveness

Error rate and error distribution

For the psychologist, mean error rate was 34.11 errors per draft note (SD = 12.33, range =18-59) (Fig. 4). The most frequent error type was "Missing positive statements" (9.11 (SD = 3.48)), while "Failure to distinguish speakers" (0.11 (SD = 0.33)) occurred least frequently (Appendix D).

For the nurse, mean error rate was 9.33 errors per note (SD = 3.50, range = 6-16) (Fig. 4). The most frequent error type was "Repetition of the same point" (2.33 (SD = 2.06)) and "Irrelevant information" did not occur in any session (Appendix D).

Association between error rate and documentation time

No significant association was observed between error rate and documentation time for the psychologist (Spearman's ρ = 0.000, p = 1.000, 95 % CI [–0.677, 0.677]) or the nurse (Pearson's r = –0.506, p = 0.305, 95 % CI [–0.934, 0.518]).

Association between error rate and session

No significant association was observed between error rate and session number for either the psychologist (Pearson's r = 0.409, p = 0.274, 95 % CI [–0.350, 0.844]) or the nurse (Pearson's r = 0.672, p = 0.144, 95 % CI [–0.308, 0.960]).

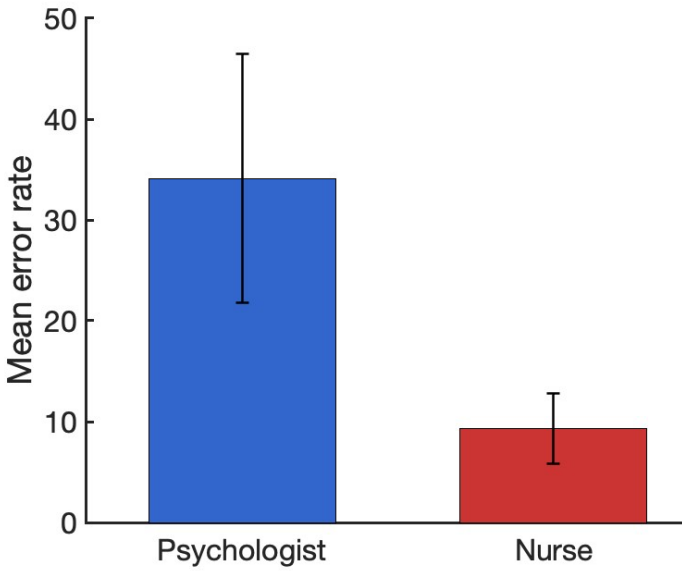


Fig. 4: *Mean total error per draft note for the psychologist (blue) and the nurse (red). Bars show mean values, and error bars indicate standard deviations (SD)*

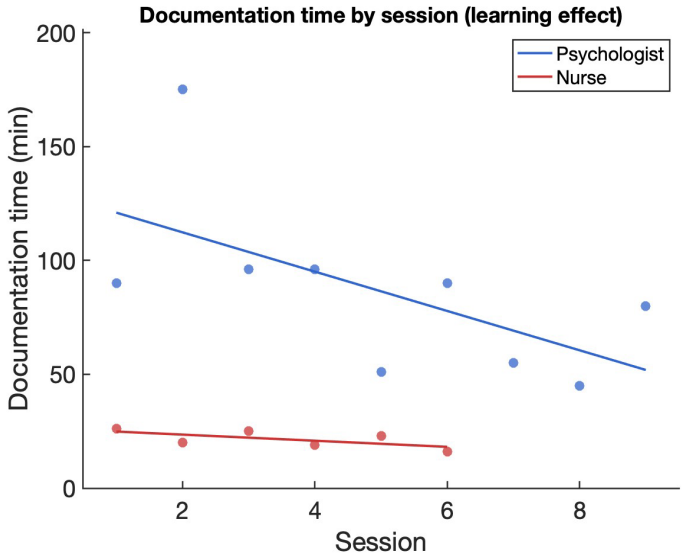


Fig 3: *Documentation time by session for the psychologist (blue) and the nurse (red). The dots represent observed documentation times, and the lines indicates the linear trend.*

Efficiency					
Analysis	**Participant**	**Baseline Mean (SD)**	**W/ Corti Mean (SD)**	**Range**	**P-value**
Documentation time with vs. without Corti (min)	Psychologist	62 (14.83) (M)	86.4 (38.82) (M)	45–175	p = .190 (M)
	Nurse	7.5 (0) (M)	21.5 (3.83) (M)	16–26	p = .024 (M) *
Effectiveness					
Error rate	Psychologist	-	34.11 (12.33)	18–59	-
	Nurse	-	9.33 (3.50)	6–16	-

Table 1: *Overview of results for documentation time (baseline vs. w/) and Error rate.*
*(M) = Mann-Whitney U-test; * = significant.*

Efficiency				
Analysis	**Participant**	**95% CI**	**Correlation Coefficient**	**P-value**
Learning effect	Psychologist	[–0.932, –0.022] (S)	ρ = –.689 (S)	p = .040 (S) *
	Nurse	[–0.958, 0.334] (P)	r = –.655 (P)	p = .158 (P)
Effectiveness				
Error rate vs. documentation time	Psychologist	[–0.677, 0.677] (S)	ρ = .000 (S)	p = 1.000 (S)
	Nurse	[–0.934, 0.518] (P)	r = –0.506 (P)	p = .305 (P)
Error rate vs. session number	Psychologist	[–0.350, 0.844] (P)	r = .409 (P)	p = .274 (P)
	Nurse	[–0.308, 0.960] (P)	r = 0.672 (P)	p = .144 (P)

Table 2: *Overview of results for correlations tests.*
*(P) = Pearson's correlation test; (S) = Spearman's rho correlation test; * = significant.*

3.2 Qualitative Results

The qualitative analysis identified five themes that described the psychologist's satisfaction using Corti for clinical documentation.

3.2.1 Satisfaction

Time Uzalition

Corti generated draft notes rapidly, which was appreciated, but this did not reduce overall documentation time. The draft notes required thorough review and editing, especially when information was missing, and therefore:

> *"…things go a bit slowly."*

Satisfaction was ambivalent, as the rapid generation of draft notes was valued, but the need for thorough review limited overall satisfaction. With increased experience and more complete draft notes, the process could have become more efficient, ideally allowing the psychologist to focus on removing irrelevant content rather than adding information:

> *"I would much rather that it [Corti] writes more, so my task becomes more of a deletion task."*

An additional consideration regarding time usage was that a single error in the error classification scheme did not necessarily reflect the time required to edit it, as it could range from a minor edit to several lines of text:

> *"I only add one mark [in the error classification scheme], but it [Corti] rarely corresponds to just a single missing line in the draft note. It's a lot."*

Workflow and Trust

Corti was considered highly usable and intuitive:

> *"It [Corti] is very intuitive…".*

Corti had not changed existing workflows, and the psychologist continued to write personal notes during consultations as an essential supplement:

> *"I don't write fewer of them [notes], because I don't trust it [Corti] enough."*

The built-in prompting feature was not used, although the psychologist saw its potential:

> *"I haven't tried prompting, but it might make it [Corti] better. Probably."*

Limited trust in Corti, due to frequent errors, contributed to the fact that Corti had not yet reduced documentation time.

Consultation Complexity

Corti performed best in one-to-one consultations with children and adolescents:

> *"It [Corti] works better when you are just sitting with one other person."*

In consultations involving multiple participants, such as consultations with parents who held different perspectives or consultations involving psychoeducation, the draft notes became less accurate:

> *"When several viewpoints are introduced, I think it [Corti] sometimes gets a bit confused."*

These complex consultations create challenges for Corti's ability to distinguish between individual speakers and to capture the professional nuances involved, such as those associated with psychoeducation.

Content, Structure and Stability

Corti sometimes failed to include both positive and negative information and basic formalities such as *"date"* or *"headings",* which required editing. More serious errors occurred when the system hallucinated:

> *"The most unfortunate is when it [Corti] completely makes something up that we haven't talked about at all."*

At the same time, Corti provided a standardised structure that made it clear to both the psychologist and colleagues where certain patient information can be found. Although errors and missing information reduced satisfaction, the standardised structure was seen as an advantage. Corti generally operated reliably and remained stable during the sessions:

> *"It [Corti] doesn't disturb."*

Some technical issues still occurred, such as recordings stopping when the computer became inactive, resulting in gaps in the draft notes. There was still room for improvement:

> *"I still think it [Corti] can be improved quite a lot."*

Overall, Corti was generally viewed positively.

Learning and Adaptation to Clinical Practice

Initially, documentation was ineffective due to the absence of guidelines on how to provide feedback to Corti. As the psychologist became familiar with the system's structure, feedback became more targeted, improving the quality of the draft notes:

> *"The clearer I was about what I wanted included, the easier it became for Corti."*

With repeated use, the psychologist began to recognise frequent error patterns:

> *"It's [Corti's] getting better […] I now have a sense of which errors will appear."*

This reflects a learning effect. Increased familiarity with the system and more precise feedback led to noticeable improvements in the draft notes. Initially, Corti had a strong medical focus that did not align with the psychologist's needs:

> *"My guess is that it [Corti] works well for physicians, for more diagnostic consultations. When I start talking about their [the patients'] experience of something […], it [Corti] becomes a bit more challenging."*

Following more targeted feedback to Corti, the medical focus was reduced, and the draft notes became more aligned with psychological practice, which enhanced satisfaction:

> *"I also think it [Corti] has removed the medical focus […] it has become a bit less."*

3.2.1.1 Summary of Satisfaction

Overall, the findings showed that the psychologist's satisfaction with Corti was ambivalent. Corti was valued for its ability to rapidly generate draft notes, its usability, technical stability, and standardised structure, all of which contributed to a more manageable documentation process. However, satisfaction was limited by errors, missing information, hallucinations, and challenges in handling multiple speakers and capturing professional nuances. The limited trust in the draft notes, combined with the fact that the number of errors did not consistently correspond to time required for editing, means that Corti, in its current form,

does not reduce documentation burden, despite ongoing adjustments based on feedback that had improved the quality of the draft notes.

4. Discussion

This study aimed to investigate how ambient AI affects documentation work in clinical practice, focusing on *efficiency*, *effectiveness* and *satisfaction*. Overall, the findings indicate ambivalent satisfaction with the use of ambient AI-scribes, which partially contrasts with previous studies reporting increased work satisfaction following the implementation of ambient AI-scribes [18,20,23–25]. Lukac et al.[18], Stults et al.[24], and Shah et al.[20] demonstrated a significant increase in work satisfaction, while this study points to a more ambivalent experience due to errors, missing information, and time-consuming post-editing. This ambivalence aligns more closely with findings reported by Shah et al.[25], where more mixed user experiences were described.

The quantitative results showed that the nurse had a significantly increased time spent on documentation with Corti. The psychologist also demonstrated increased documentation time, although this was not statistically significant. These findings differ from previous studies [18,21,22,24,26,27], in which ambient AI is associated with reduced documentation time.

One possible explanation relates to the context of use in this study. The nurse's consultations were conducted by telephone, which may challenge Corti's ability to accurately capture spoken content during the consultation. This may result in incomplete draft notes and time-consuming post-editing. According to the psychologist, another challenge was present in the psychologist's consultations involving multiple participants. The increased consultation complexity places high demands on Corti's ability to capture spoken content and distinguish between individual speakers, a challenge that has also been reported in studies of similar technologies [32,33]. In both cases, Corti's ability to generate comprehensive and accurate draft notes is challenged. However, despite the psychologist's perception that distinguishing between speakers constituted a major challenge, quantitative findings showed that this error type was the least frequent in the psychologist's consultations. This may indicate that the complexity of multi-participant consultations heightens the clinician's awareness of potential errors.

The psychologist's experience of insufficient draft notes was supported by quantitative findings, with missing positive information being the most frequent error type. This aligns with findings from previous studies reporting frequent omission of information and the need for manual post-editing in similar systems [34,35].

This may be explained by Corti's use of the FactsR agent [36], which is designed to filter information and generate short, medically focused draft notes.

Guo et al.[21], Shah et al.[25], and Stults et al.[24] reported that ambient AI-scribes generate longer draft notes than manually written notes. However, none of the studies explicitly mention functionality similar to the FactsR agent, making it uncertain whether comparable features were used. Differences in system design may therefore contribute to variation in note length. An alternative explanation may be differences in professional documentation needs, as the cited studies mainly involve physicians, who may prefer short and diagnostic draft notes, while psychologists may require more detailed descriptions.

Although this study does not demonstrate a time increase for either clinician, this

is not unusual. A scoping review by Tsai et al.[37] shows that new EHR initiatives often lead to an initial increase in time use, with time savings emerging later. This is supported by the quantitative findings of this study, where both clinicians demonstrated a positive learning effect over time, although this was only statistically significant for the psychologist. This indicates
that increased familiarity with Corti may potentially reduce documentation time in the long term.

A positive but non-significant association between error rate and session number was observed for both clinicians. The quantitative results showed a tendency towards increasing error rates over time, which contrasts with the psychologist's qualitative experience that errors appeared to decrease. This lack of alignment may be explained by a learning effect, whereby errors are experienced as less challenging over time and therefore not reflected in a reduced error rate. This interpretation is supported by general learning effect theories described by Waldman et al. [38], suggesting that efficiency increases as individuals repeatedly perform the same task over time. The psychologist reported becoming faster at identifying and managing errors with increased experience. Finally, methodological factors may have influenced error registrations. It cannot be ruled out that the error classification scheme may have been applied less consistently throughout the study period, or that feedback to Corti could have influenced the types and frequency of errors.

The findings indicate that ambient AI scribes are likely to remain a developing technology in the near future. The significance for clinical documentation work is not yet clearly defined. In clinical practice, the findings suggest that ambient AI scribes cannot be used independently of existing documentation practices. Future research should prioritize longitudinal studies with larger sample sizes across multiple clinical contexts to evaluate how ambient AI affects clinical documentation work over time.

4.1 Limitations

This study has limitations, including the possibility that the error rate may have been influenced by variation in how participants applied the error classification scheme. Because the error classification scheme was not standardised, it may have limited measurement validity, while its use by multiple individuals and potential inconsistency over time may have affected inter- and intrarater reliability. During data collection, participants also spent time entering data into the scheme, which introduces a bias in relation to the efficiency outcomes.

This study is further limited by a small sample size, restricted data, and a short data-collection period, why these findings should be interpreted cautiously. The qualitative interview was conducted only with the psychologist, as the nurse joined the study later. Consequently, due to the limited time frame of the study, it was not possible to include an interview with the nurse, and her satisfaction could not be evaluated.

Furthermore, this study does not fully align with a traditional explanatory sequential mixed-methods design, as only six of the nine quantitative observations had been collected from the psychologist before the qualitative phase commenced.

5. Conclusion

The implementation of Corti did not result in increased efficiency and was associated

with ambivalent user satisfaction. Effectiveness was not perceived as sufficient at the current stage of implementation.

These findings should be interpreted cautiously due to study limitations. Future research should therefore prioritize longitudinal studies across multiple clinical settings to evaluate the impact of ambient AI on clinical documentation work.

6. References

1. Nguyen MLT, Honcharov V, Ballard D, Satterwhite S, McDermott AM, Sarkar U. Primary Care Physicians' Experiences With and Adaptations to Time Constraints. JAMA Netw Open. 30. april 2024;e248827.
2. Tajirian T, Stergiopoulos V, Strudwick G, Sequeira L, Sanches M, Kemp J, m.fl. The Influence of Electronic Health Record Use on Physician Burnout: Cross-Sectional Survey. J Med Internet Res. 15. juli 2020;22(7):e19274.
3. Gaffney A, Woolhandler S, Cai C, Bor D, Himmelstein J, McCormick D, m.fl. Medical Documentation Burden Among US Office-Based Physicians in 2019: A National Study. JAMA Intern Med. 1. maj 2022;182(5):564–6.
4. Shanafelt TD, Dyrbye LN, Sinsky C, Hasan O, Satele D, Sloan J, m.fl. Relationship Between Clerical Burden and Characteristics of the Electronic Environment With Physician Burnout and Professional Satisfaction. Mayo Clin Proc. juli 2016;91(7):836–48.
5. Micek MA, Arndt B, Tuan WJ, Trowbridge B, Dean SM, Lochner J, m.fl. Physician Burnout and Timing of Electronic Health Record Use. ACI Open. januar 2020;4(1):e1–8.
6. Adler-Milstein J, Zhao W, Willard-Grace R, Knox M, Grumbach K. Electronic health records and burnout: Time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. J Am Med Inform Assoc JAMIA. 1. april 2020;27(4):531–8.
7. Apathy NC, Rotenstein L, Bates DW, Holmgren AJ. Documentation dynamics: Note composition, burden, and physician efficiency. Health Serv Res. 2023;58(3):674–85.
8. Gardner RL, Cooper E, Haskell J, Harris DA, Poplau S, Kroth PJ, m.fl. Physician stress and burnout: the impact of health information technology. J Am Med Inform Assoc JAMIA. 1. februar 2019;26(2):106–14.
9. Li C, Parpia C, Sriharan A, Keefe DT. Electronic medical record-related burnout in healthcare providers: a scoping review of outcomes and interventions. BMJ Open. 19. august 2022;12(8):e060865.
10. Linzer M, Smith CD, Hingle S, Poplau S, Miranda R, Freese R, m.fl. Evaluation of Work Satisfaction, Stress, and Burnout Among US Internal Medicine Physicians and Trainees. JAMA Netw Open. 1. oktober 2020;3(10):e2018758.
11. Babbott S, Manwell LB, Brown R, Montague E, Williams E, Schwartz M, m.fl. Electronic medical records and physician stress in primary care: results from the MEMO Study. J Am Med Inform Assoc JAMIA. februar 2014;21(e1):e100-106.
12. Gesner E, Dykes PC, Zhang L, Gazarian P. Documentation Burden in Nursing and Its Role in Clinician Burnout Syndrome. Appl Clin Inform. 19. oktober 2022;13:983–90.
13. Olivares Bøgeskov B, Grimshaw-Aagaard SLS. Essential task or meaningless burden? Nurses' perceptions of the value of documentation. Nord J Nurs Res. 1. marts 2019;39:9–19.
14. Topaz M, Peltonen LM, Zhang Z. Beyond human ears: navigating the uncharted risks of AI scribes in clinical practice. Npj Digit Med. 24. september 2025;8:569.
15. Wendt SJ, Dinh CT, Sutcliffe M, Jones K, Scanlan JM, Smitherman JS. Deploying ambient clinical intelligence to improve care: A research article assessing the impact of nuance DAX on documentation burden and burnout. Future Healthc J. 17. juli 2025;12:6.
16. Sarraf B, Ghasempour A. Impact of artificial intelligence on electronic health record-related burnouts among healthcare professionals: systematic review. Front Public Health. 3. juli 2025;13:13.
17. Falcetta FS, De Almeida FK, Lemos JCS, Goldim JR, Da Costa CA. Automatic documentation of professional health interactions: A systematic review. Artif Intell Med. marts 2023;137:16.

18. Lukac PJ, Turner W, Vangala S, Chin AT, Khalili J, Shih YCT, m.fl. A Randomized-Clinical Trial of Two Ambient Artificial Intelligence Scribes: Measuring Documentation Efficiency and Physician Burnout. [henvist 28. oktober 2025]; Tilgængelig hos: https://www.medrxiv.org/content/10.1101/2025.07.10.25331333v1
19. Patterson J, Kovacs M, Lees C. Ambient Artificial Intelligence Scribes: A Pilot Survey of Perspectives on the Utility and Documentation Burden in Palliative Medicine. Healthcare. 26. august 2025;13:2118.
20. Shah SJ, Devon-Sand A, Ma SP, Jeong Y, Crowell T, Smith M, m.fl. Ambient artificial intelligence scribes: physician burnout and perspectives on usability and documentation burden. J Am Med Inform Assoc JAMIA. 5. december 2024;32(2):375–80.
21. Guo Y, Hu D, Wang J, Zheng K, Perret D, Pandita D, m.fl. Ambient Listening in Clinical Practice: Evaluating EPIC Signal Data Before and After Implementation and Its Impact on Physician Workload. I: Househ MS, Tariq ZUA, Al-Zubaidi M, Shah U, Huesing E, redaktører. Studies in Health Technology and Informatics [Internet]. IOS Press; 2025 [henvist 28. oktober 2025]. Tilgængelig hos: https://ebooks.iospress.nl/doi/10.3233/SHTI250921
22. Ma SP, Liang AS, Shah SJ, Smith M, Jeong Y, Devon-Sand A, m.fl. Ambient artificial intelligence scribes: utilization and impact on documentation time. J Am Med Inform Assoc. 1. februar 2025;32:381–5.
23. Albrecht M, Shanks D, Shah T, Hudson T, Thompson J, Filardi T, m.fl. Enhancing clinical documentation with ambient artificial intelligence: a quality improvement survey assessing clinician perspectives on work burden, burnout, and job satisfaction. JAMIA Open. 26. december 2024;8:ooaf013.
24. Stults CD, Deng S, Martinez MC, Wilcox J, Szwerinski N, Chen KH, m.fl. Evaluation of an Ambient Artificial Intelligence Documentation Platform for Clinicians. JAMA Netw Open. 2. maj 2025;8(5):13.
25. Shah SJ, Crowell T, Jeong Y, Devon-Sand A, Smith M, Yang B, m.fl. Physician Perspectives on Ambient AI Scribes. JAMA Netw Open. 24. marts 2025;(3):9.
26. Olson KD, Meeker D, Troup M, Barker TD, Nguyen VH, Manders JB, m.fl. Use of Ambient AI Scribes to Reduce Administrative Burden and Professional Burnout. JAMA Netw Open. 2. oktober 2025;(10):12.
27. Rodenhouse AJ, Davis E, Burger P, Nicandri G, DiGiovanni B. Utilization of an Artificial Intelligence–Based Documentation System Improves Provider Efficiency in Outpatient Orthopaedic Clinics: Reducing the Afterhours Burden of the Electronic Health Record—A Pilot Study. J Am Acad Orthop Surg. 16. september 2025;28.
28. Aalborg Universitetshospital. aalborguh.rn.dk. [henvist 4. december 2025]. Steno Diabetes Center Nordjylland. Tilgængelig hos: https://aalborguh.rn.dk/afsnit-og-ambulatorier/endokrinologisk-afdeling/saerlige-funktioner/steno-diabetes-center-nordjylland
29. Appel U, Larsen HR. Talegenkendelse. Projektbeskrivelse. Steno Diabetes Cent Nord. 15. oktober 2025;(2.0):12.
30. Dansk Standard. Ergonomi – Interaktion mellem menneske og system – Del 11: Brugbarhed: Definitioner og begreber [Internet]. Nordhavn, Danmark: Danish Standards Association; 2018 jun [henvist 28. oktober 2025] s. 42. Report No.: DS/EN ISO 9241-11:2018. Tilgængelig hos: https://www.ds.dk
31. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol. 1. januar 2006;3:77–101.
32. Quiroz JC, Laranjo L, Kocaballi AB, Berkovsky S, Rezazadegan D, Coiera E. Challenges of developing a digital scribe to reduce clinical documentation burden. Npj Digit Med. 22. november 2019;2(1):114.
33. Mess SA, Mackey AJ, Yarowsky DE. Artificial Intelligence Scribe and Large Language

Model Technology in Healthcare Documentation: Advantages, Limitations, and Recommendations. Plast Reconstr Surg Glob Open. 16. januar 2025;13(1):e6450.
34. Biro J, Handley JL, Cobb NK, Kottamasu V, Collins J, Krevat S, m.fl. Accuracy and Safety of AI-Enabled Scribe Technology: Instrument Validation Study. J Med Internet Res. 27. januar 2025;27:e64993.
35. Kernberg A, Gold JA, Mohan V. Using ChatGPT-4 to Create Structured Medical Notes From Audio Recordings of Physician-Patient Encounters: Comparative Study. J Med Internet Res. 22. april 2024;26:e54419.
36. Corti. Corti API Documentation. [henvist 28. oktober 2025]. Introducing FactsR™. Tilgængelig hos: https://docs.corti.ai/about/factsr
37. Tsai CH, Eghdam A, Davoody N, Wright G, Flowerday S, Koch S. Effects of Electronic Health Record Implementation and Barriers to Adoption and Use: A Scoping Review and Qualitative Analysis of the Content. Life. 4. december 2020;10(12):327.
38. Waldman JD, Yourstone SA, Smith HL. Learning Curves in Health Care. Health Care Manage Rev. 2003;28(1):41–54.

Appendix list

Appendix A: Error Classification Scheme

Error classification scheme
Date:
Conversation complexity (low, medium, high):
Time spent correcting the Corti note:

Rephrasing / Corrections	
Words	
Sentence(s)	
Entire paragraph(s)	
Missing information	
Lack of positive or non-problematic statements (e.g., related to sleep or diet) that are still important information	
Lack of negative or problematic statements (e.g., related to sleep or diet) that are still important information	
Missing formalities and factual details (date, participants, headings, sections, etc.)	
Misinterpretations and false conclusions	
Factual errors (e.g., incorrect weekday) and hallucinations (e.g., the patient has never been hospitalized)	
Concluding / Diagnostic	
Errors in language	
Sentences without meaning	
Repetition of the same point	
Who said what	
Does not distinguish between individuals (e.g., mixes up different statements)	
Relevance	
Irrelevant information, e.g., an overemphasis on equipment/technology instead of psychological aspects	
Other	

Appendix B: Interview guide

Briefing

”Thank you for agreeing to participate in this interview.
The purpose of the interview is to gain insight into your experiences with using Corti in your work. We are interested in hearing about both the positive experiences and any challenges you may have encountered in connection with the use of the system.

If, during the interview, we ask a question that you cannot immediately answer, it is completely okay to take time to think about the answer. If you still do not know what to answer, it is better that you say so rather than make up an answer. It is important that you feel comfortable with the situation and answer the questions honestly. Before we move on to a new question during the interview, we will try to ask whether you have anything to add before moving on to the next question. If at any point during the interview you feel that you are left with something important, please feel free to say so.

Before we begin, we would like to ask for your consent for the interview to be recorded. We use a mobile phone as well as a voice recorder to record the conversation. The recording is used exclusively to ensure accuracy when your answers are subsequently reviewed and analyzed. I may note a few keywords during the interview to remember to follow up on some of your statements. All information is treated confidentially and anonymized before being included in the project.”

The consent form is signed.

“Is it okay that we begin?”

Theme 1: Satisfaction

Research question	Interview questions
How is the use of Corti experienced and how is the experience of satisfaction affected?	1. How would you describe your overall experience of working with Corti? 2. Which aspects of Corti do you experience as the most helpful or valuable? 3. In which situations do you experience frustrations or dissatisfaction when using Corti? 4. How does Corti affect your experience of presence and contact in conversations with patients and/or relatives? 5. How has Corti affected your motivation or job satisfaction in everyday work? (e.g. experience of calm, workflow, or documentation burden) 6. If you were to describe your ideal version of Corti, what would it be able to do differently?

Theme 2: Effectiveness

Research question	Interview questions
How is the use of Corti experienced in relation to the quality of the technology's output?	7. How do you experience the quality of what Corti produces in relation to your professional standards? 8. Which types of errors do you experience as the most serious? 9. Which errors occur most frequently? (Descriptive data) 10. How do you handle errors when you discover them, and how does this affect your work? 11. How do you assess Corti's ability to capture nuances and context in conversations? 12. How do you experience that the frequency of errors changes over time – that is, from session to session? (Session vs. errors) 13. How do any errors affect your trust in Corti?

Theme 3: Efficiency

Research question	Interview questions
How is the use of Corti experienced in relation to efficiency and time consumption?	14. What significance has Corti had for your experience of efficiency in everyday work? 15. How do you experience that Corti affects your daily time consumption on documentation? – Is it better with or without Corti? – What significance do you experience that it has for the number of consultations you can complete? 16. In which situations (if any) do you experience that Corti takes extra time? – How do you experience the relationship between the number of errors and the time you spend correcting/documenting? 17. In which situations (if any) do you experience that Corti has made the work faster for you? 18. How would you describe your development in relation to the time you spend documenting with Corti, as you do it more times?

Theme 4: Workflows

Research question	Interview questions

How is the use of Corti experienced to affect daily workflows and the organization of work?	19. How does Corti fit into your daily practice and workflows? 20. Which types of tasks or conversations do you experience as being best suited for the use of Corti? – Does the number of people in the room matter? 21. When you think about your consultations before and after Corti, which differences do you notice in your approach or preparation? 22. Which adjustments have you made to your way of working to make Corti function as well as possible? 23. How do you see Corti's role in your work going forward, as support, relief, or something else?

Debriefing

"These were the questions we had.

Before we finish, we would like to ask if there is anything you would like to add. Is there anything we have not talked about but that you think is important in relation to Corti and your work?

Thank you for taking the time to participate and share your experiences. Your responses will be an important part of the analysis in the project."

Appendix C: Transcription guidelines

Speaker identification
Speakers are consistently marked using fixed initials:
I1 = Interviewer 1
I2 = Interviewer 2
D = Participant
Example:
I1: How did you experience the process?
D: I felt well prepared.

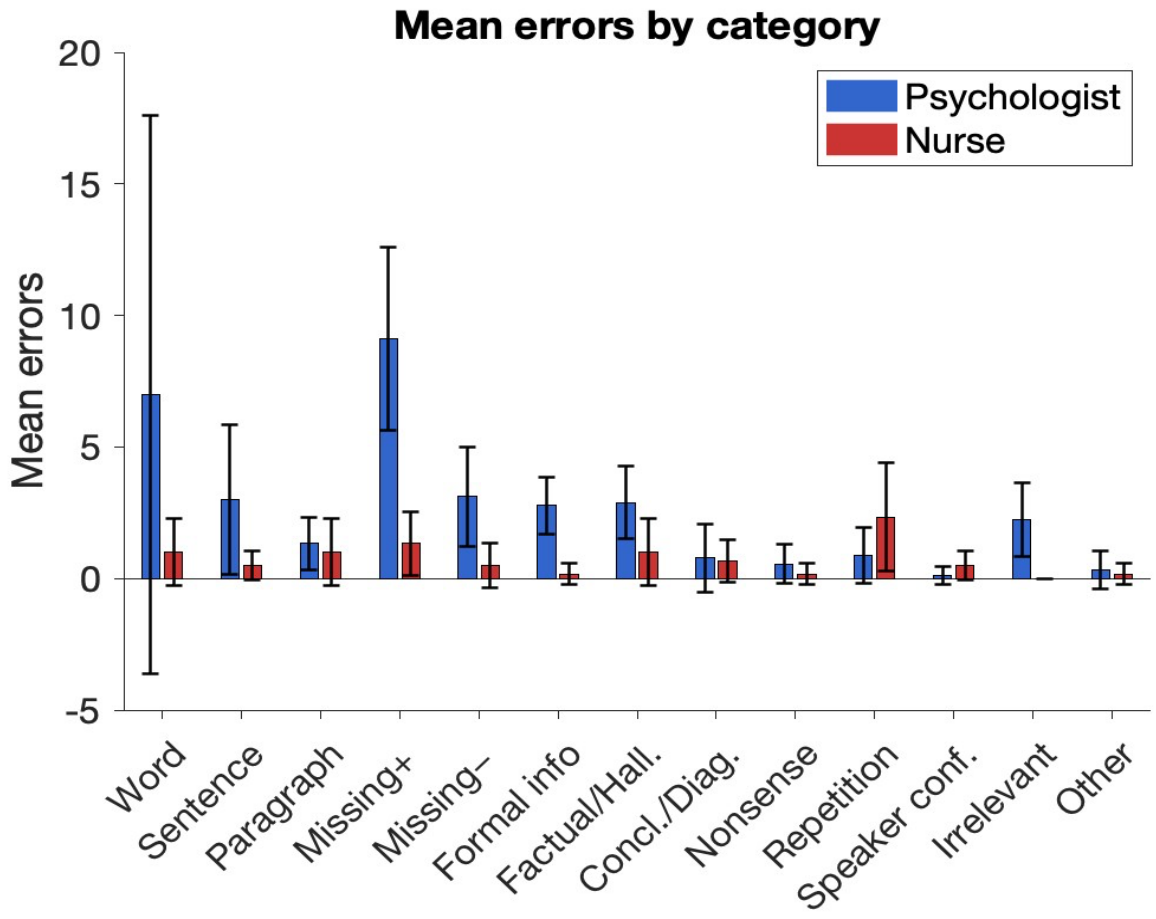
Line breaks when speakers change
Each time a new person speaks, the statement is written on a new line.

Timestamps
Timestamps are added every 2 minutes or at key transitions.
Format: [00:02:00]
Example:
D: It was challenging at the beginning. [00:02:00]

Element	**Rule**	**Example of how it should appear in the transcript**
Short pause (under 2 sec.)	Not noted	—
Long pause (over 2 sec.)	Marked with ...	*Hmm ... that is a good question.*
When the participant interrupts themself / changes direction mid-sentence	Marked with (...)	Spoken: *It was okay, I think. I feel it was very difficult.* → Transcribed: *It was okay (...) I feel it was very difficult.*
Filler words (uh, well, mmh)	Removed	Spoken: *So uh I think uh it was fine* → Transcribed: *So I think it was fine.*
Repetitions without meaning	Removed	Spoken: *it was… it was… it was actually easy → It was actually easy.*
Repetitions with meaning	Kept	*I was really, really stressed.*
Hesitation, irony, laughter, sighs, emotional reactions, etc.	Marked in parentheses	*I was (hesitates) actually unsure.*
Unclear word	Noted as [unclear]	*I think [unclear] was the reason.*

Guessing a word	Marked with a question mark	*We had that kind of [registration?] in the system.*
Inaudible sequence	Marked with [...]	[...]
Interruption	Marked with (interrupts)	Participant: *I experienced—* Interviewer: *(interrupts) Can you explain that?*

Appendix D: Distribution of mean errors per category



Bar plots showing the distribution of mean errors per category for the psychologist and the nurse Bars represent mean values and error bars indicate standard deviations (SD). Categories: Word = word reformulation; Sentence = sentence reformulation; Paragraph = paragraph reformulation; Missing+ = missing positive statements; Missing− = missing negative statements; Formal info = missing formal information; Factual/Hall. = factual errors or hallucinations; Concl./Diag. = conclusive or diagnostic statements; Nonsense = nonsensical sentences; Repetition = repetition of the same point; Speaker conf. = failure to distinguish speakers; Irrelevant = irrelevant information; Other = other error