

Accuracy and Safety of an AI Ambient Scribe Compared with Handwritten Clinical Notes

Byron De John

byron.de.john@uct.ac.za

University of Cape Town

Johannes M.N Enslin

University of Cape Town

Joshua Fieggen

University of Oxford

Linda Camara

GraiLabs, Cape Town

Bruce Bassett

University of Witwaterstrand

Graham Fieggen

University of Cape Town

Article

Keywords:

Posted Date: April 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9139641/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: Competing interest reported. Linda Camara is employed by GraiLabs (Head of Product Development) and developed the ambient AI scribe system evaluated in this study, providing technical support for system configuration and deployment only and was not involved in study design, outcome assessment, statistical analysis, interpretation, or manuscript drafting. All other authors declare no competing interests.

Abstract

In a prospective real-world evaluation at Groote Schuur Hospital, South Africa, a large-language-model ambient scribe was compared with contemporaneous handwritten clinical notes. The system generated notes from raw audio without diarisation, transcript editing, or clinician review. Across 49 encounters, documentation quality was independently assessed using a SOAP-aligned rubric (0–5 per domain) and a symmetric severity-graded error taxonomy. AI-generated notes outperformed handwritten notes in 48 encounters and tied in one, with higher mean overall SOAP scores (4.9 vs 2.9) and a 97.1% posterior probability (95% credible interval, 91.0%–99.8%) of superior documentation quality. Posterior rates of moderate-to-severe hallucinations, distortions, omissions, and clinically significant errors were at least fivefold higher in handwritten notes. Hallucinations were not confined to AI outputs, challenging their framing as an AI-specific risk. In LMIC settings, ambient AI scribes could complement existing documentation workflows and may form part of a broader pathway toward scalable digital health infrastructure.

1. Introduction

Although Artificial intelligence (AI) may advance healthcare through improved diagnosis, treatment personalisation, administrative efficiency, and drug discovery, challenges related to data quality, regulation, and clinical integration persist.[1–4] Clinical documentation is fundamental to patient safety, continuity of care, and medico-legal accountability. Large language models (LLMs) have the potential to reduce documentation burden, while supporting diagnostic reasoning and patient communication, however real-world evidence remains sparse, and previous work has predominantly focused on high-income settings.[5, 6]

In many low- and middle-income countries (LMICs) documentation relies on handwritten (HW) notes that vary in completeness, legibility, and structure. These records require substantial infrastructure and personnel for storage, retrieval, and distribution. Ambient AI scribes that transform real-time audio into structured clinical notes have emerged as a potential means to reduce administrative burden while maintaining, or potentially improving, accuracy and medico-legal adequacy.[7]

To our knowledge, no prior study has directly compared the quality, completeness, and error profiles of clinician-generated HW notes with those produced by an ambient, LLM-based AI scribe in a LMIC setting. Most evaluations of AI documentation systems have benchmarked model performance against audio transcripts, automated metrics, or user-experience surveys. [8–15] We therefore designed a prospective, real-world proof-of-concept evaluation of an ambient AI scribe in an LMIC setting, using predefined scoring domains aligned with accepted clinical note structures and a symmetric error taxonomy applied equally to AI-generated and handwritten notes.

2. Results

2.1 Cohort and data completeness

A total of 53 clinical encounters were recorded. One encounter was used exclusively for prompt optimisation and was prospectively excluded from analysis. Two encounters were excluded due to early device-related truncation errors rendering them unscorable. A further encounter was excluded because it contained insufficient content.

The final dataset comprised 49 encounters spanning the full neurosurgical service, with no patients declining recording.

2.2 Encounter Characteristics and Performance Metrics

Encounter characteristics, documentation efficiency, content density and comparative SOAP rubric scores for AI and HW notes are summarised in Table 1. The AI’s SOAP score performance varied only minimally across the audio recording quality strata.

Table 1. Encounter characteristics and comparative documentation metrics for AI-generated and handwritten notes (n = 49)

	Handwritten notes	AI-generated notes
Encounter characteristics		
Total encounters	49	
Location of recording	Outpatient department: 21 (42.9%) Neurosurgical ward: 12 (24.5%) Operating theatre: 4 (8.2%) Trauma unit: 3 (6.1%) Medical emergency unit: 3 (6.1%) Non-neurosurgical wards: 3 (6.1%) Unspecified*: 3 (6.1%)	
Note type	Consultation: 45 (91.8%) Operative: 4 (8.2%)	
Clinician seniority	Registrar: 39 (79.6%) Consultant: 10 (20.4%)	
Mean recording length, min (range)	12.3 (2.1–42.7)	
Mean transcript word count (range)	1 577 (222–4 317)	
Audio quality	Good: 39 (79.6%) Satisfactory: 8 (16.3%) Poor: 2 (4.1%)	
Handwritten note properties		
Legibility	Legible: 15 (30.6%) Partially legible: 25 (51.0%) Illegible: 9 (18.4%)	—
Mean abbreviations per note (range)	12.4 (0–36)	—
Medico-legal completeness	Signature: 44 (89.8%) Date: 44 (89.8%) Time: 43 (87.8%) Location: 41 (83.7%) Legible clinician name: 27 (55.1%)	—
Comparative documentation metrics		
Mean word count (range)	116 (20–377)	808 (332–1 446)
Mean time to create note, min (range)	7.9 (1.2–26.0)	2.1 (0.75–5.6)
Mean total SOAP score (0–5)	2.9	4.9
Wins by overall SOAP score	0 (0.0%)	48 (98.0%)
Mean SOAP rubric scores by clinician seniority		
SOAP domain	Consultant Registrar	AI
Subjective	3.7 2.7	4.9

Objective	3.9 2.6	4.8
Assessment	2.9 2.3	4.9
Plan	3.8 2.9	4.9
Overall organisation	4.1 3.0	4.9
Overall mean SOAP score	3.7 2.7	4.9
Mean time to make note, min	8.2 7.8	2.1
ICD-10 coding (AI only)†		
Correct primary and secondary coding	N/A	94.0%
Complete coding (all secondary codes)	N/A	98.0%

Abbreviations: HW = handwritten; SOAP = Subjective, Objective, Assessment, Plan; SOAP scores on a 0–5 Likert scale.

* Clinician did not indicate location in the HW note or audio recording.

† HW notes did not include ICD-10 coding in routine practice; AI coding is reported as a unilateral assessment.

In the SOAP breakdown, HW columns show Consultant | Registrar scores. One encounter resulted in a tie.

2.3 Documentation quality

In head-to-head comparisons, AI-generated notes consistently outperformed HW documentation on overall SOAP rubric scores (Table 1 and Figure 1). The estimated probability that the AI system produces a higher-quality note than the HW method in a typical encounter was 97.1% (95% credible interval, 91.0%–99.8%), whereas the inverse probability was negligible (7×10^{-15}). Figure 1 plots overall individual SOAP rubric scores against time taken to complete the note. For HW notes, higher overall rubric scores were associated with longer documentation times.

Figure 1: Graph of individual overall rubric SOAP scores vs time to make notes

2.4 Error profile and severity

Error distributions between AI-scribe and handwritten notes are presented in figure 2 and Supplemental Table 1. Distortions were uncommon, but occurred more frequently in HW notes than in AI notes, particularly at higher severity (AI, n = 5; HW, n = 15). Grade 4 distortions occurred predominantly in HW notes, corresponding to an estimated fivefold higher distortion rate for HW documentation at this

severity (median relative risk [RR], 5.22; 95% credible interval [CrI], 0.98–61.9; posterior probability that $HW > AI$, 0.97). Lower-grade distortions were similar between groups.

Hallucinations were observed at similar frequencies, but their severity distributions differed substantially (AI, $n = 21$; HW, $n = 23$). AI hallucinations were mostly minor (grades 1–2), whereas HW demonstrated a predominance of clinically meaningful hallucinations. At severity grade 3, HW documentation was associated with a markedly higher hallucination rate (median RR, 16.2; 95% CrI, 3.76–184.0; posterior probability = 1.0). Omissions represented the dominant error mode in HW documentation and were more frequent and more severe than in AI notes (AI $n = 15$; HW $n = 131$). Across severity grades 1–4, HW notes exhibited consistently higher omission rates, with median relative risks ranging from 5.2 to 154.7 and posterior probabilities approaching 1.0 across all clinically relevant severities. Notably, grade 4 omissions occurred only in HW notes. Specific examples of each error type are provided in Addendum A2.

A subset of AI-related errors was attributable to transcription failure rather than content generation. All AI distortions arose from transcription errors (5 of 5, 100.0%), as did a minority of hallucinations (2 of 21, 9.5%) and approximately two thirds of omissions (10 of 15, 66.7%). The majority of transcription-related errors were graded as low severity.

AI notes were over six times more likely to be free of clinically meaningful errors than HW notes (AI, $n = 25$; HW, $n = 4$; Figure 2). Major clinical impact was observed in 38.8% of HW notes ($n = 19$) compared with 2.0% of AI-generated notes ($n = 1$), corresponding to a markedly reduced risk of major error with AI documentation (RR, 0.06; 95% CrI, 0.006–0.25). Minor clinical impact occurred at similar frequencies between methods, affecting 53.1% of AI-generated notes ($n = 26$) and 46.9% of HW notes ($n = 23$).

Figure 2. Error profiles and clinical impact of AI-generated versus handwritten clinical documentation. (A) Raw error counts by type (distortion, hallucination, omission) and severity grade (1 = minor, 5 = catastrophic) for AI-generated (blue) and handwritten (red) notes across 49 encounters. Omissions were the dominant error mode in handwritten documentation; AI hallucinations were predominantly low-grade (grades 1–2), whereas handwritten hallucinations clustered at grade 3. (B) Bayesian relative risk estimates (HW/AI) with 95% credible intervals on a log scale. Values greater than 1 indicate higher error rates in handwritten notes. Arrows denote truncated credible intervals. (C) Overall clinical impact classification per encounter. Major documentation errors occurred in 38.8% of handwritten notes compared with 2.0% of AI-generated notes (relative risk 16.1; 95% CrI 4.0–175.4; posterior probability 1.0). Error definitions and severity grading are described in Section 2.8. CrI = credible interval; HW = handwritten.

2.5 Inter-rater reliability

Inter-rater agreement is presented in figure 3. Overall paired SOAP rubric ratings across both AI and HW scores from the two reviewers, inter-rater agreement was high: Quadratic weighted Cohen's κ was 0.814,

indicating strong agreement on the ordinal 0–5 scale. The reviewers assigned identical scores in 59% of ratings and were within one rubric point in 94.4% of cases.

Figure 3: Cross-tabulation of raw counts of rater scores showing good concordance of scores between the two raters. In 94% of cases the rater scores differed by one point or less.

3. Discussion

Current evaluations of ambient AI scribes have been conducted almost exclusively in high-income settings with established electronic health record infrastructure, focusing on transcription accuracy, text-generation quality, efficiency outcomes such as after-hours documentation burden, and clinician satisfaction.[8–15, 19, 20] None have included structured, head-to-head comparisons with handwritten clinical documentation using a shared rubric and symmetric error taxonomy – the documentation modality that predominates across LMICs.[8, 19] This study addresses these gaps. To our knowledge, it represents the first real-world comparison of handwritten documentation versus ambient AI scribe outputs in any setting, and one of the first evaluation of an ambient AI scribe conducted in a LMIC, where the documentation challenges, infrastructure constraints, and potential advantages differ substantively from those in high-income health systems.

This evaluation assessed clerical documentation performance rather than the quality of clinical management. Core elements of care were frequently articulated in appropriate detail during the clinical encounter and captured in the audio recordings, but incompletely reflected in HW documentation. The study conditions constrained AI performance in several respects: the scribe operated without speaker diarisation, transcript editing, or human correction, and clinicians were blinded to outputs. However, the design also incorporated features that may have favoured AI documentation, including a verbalisation instruction that enriched the audio transcript beyond typical clinical dialogue and the use of encounter audio as both the AI's primary input and the scoring reference standard. The net direction of these competing biases is uncertain, and the results should be interpreted accordingly.

Across nearly all encounters, the AI scribe produced documentation of higher completeness, structure, and accuracy than HW notes, with a performance advantage consistent across clinician experience. AI-generated notes were substantially more information-dense and demonstrated high reliability in diagnostic coding (ICD-10), with 94% accuracy and 98% completeness of coding across encounters (a unilateral performance assessment, as ICD-10 coding is not part of routine handwritten documentation). These findings align with prior reports that LLM-based documentation systems frequently capture more clinically relevant information and are not subject to fatigue, cognitive overload, or time pressure. Furthermore, these findings extend previous evaluations – which benchmarked AI notes against audio transcripts alone – by demonstrating superiority in direct comparison with contemporaneous clinician-written notes.[8, 19] These findings are particularly salient in LMIC settings, where HW clinical notes are frequently reported to be incomplete, poorly structured, and variably legible, with potential implications for patient safety, continuity of care, and auditability.[21, 22]

HW documentation was associated with a higher overall burden of error and a disproportionate concentration of high-severity errors, whereas errors in AI-generated notes were predominantly low grade. This difference was not confined to a single error category but was observed consistently across error subtypes and severity strata.

A notable finding was that content not supported by the encounter audio - classified as hallucination under our taxonomy - was not confined to AI-generated documentation, but occurred frequently, and at greater clinical severity, in handwritten notes. This challenges the assumption that hallucination represents a uniquely AI-associated risk.[4, 19, 20] However, because clinicians were not accustomed to working with ambient scribes, some documented findings classified as unsupported (e.g. pupillary responses) may reflect actions that were performed but not verbalised. The audio recording is therefore an imperfect reference standard for handwritten notes, in a way that it is not for AI notes, which are generated exclusively from that audio. Other handwritten hallucinations, such as consent discussions or risk counselling not evidenced in the recording, are less readily explained by unverballed actions and more plausibly reflect cognitive heuristics or template-driven documentation habits. The term "hallucination" thus carries different mechanistic implications in each modality, and direct comparison of rates should be interpreted with this asymmetry in mind.

The AI system exhibited a more favourable error profile overall, with errors less frequent, lower severity, and rarely associated with major clinical implications. A proportion of AI-related errors were attributable to upstream transcription failures, suggesting that further improvements in audio capture, diarisation, and transcription fidelity may yield additional safety gains. These findings hold particular relevance for LMIC settings, where reliance on HW documentation remains widespread due to limited uptake of EHRs. Recent reviews of AI in medicine highlight the opportunity for LMICs to "leapfrog" legacy EHR infrastructure and adopt modern, AI-enabled digital documentation systems.[2] Ambient AI scribes could therefore represent not only a documentation tool but also an enabling digital foundation for structured data capture, quality improvement, analytics, and future data-driven hospital optimisation. More broadly, these results suggest that in settings where handwritten documentation remains the default, the relevant comparator for ambient AI scribes is not a high-income setting defined gold standard, but the error-prone, and incomplete documentation that arises under real-world constraints. In addition, involving LMICs in AI-development is essential to ensure models are trained on representative populations in diverse settings, preventing the amplification of bias and inequity while enabling scalable solutions for areas with the greatest unmet clinical need.[23]

Strengths of this evaluation include real-world data collection, independent scoring and domain-aligned evaluation using a symmetric error taxonomy. As a single-centre neurosurgery evaluation of one device and one ambient scribe system in English-speaking adults, findings primarily inform feasibility and workflow performance and require multi-site, multi-specialty replication. Limitations include the relatively small sample size, restriction to English-speaking adults, reliance on a single device, and the labour-intensive nature of human scoring—an issue acknowledged in ambient scribe evaluation literature.[8] Raters were not blinded to note type because source-identifying features were intrinsic to the notes,

introducing potential assessment bias, although the high concordance between independent raters provides some reassurance. Clinicians were encouraged to verbalise findings and decisions that might otherwise be recorded only in writing, which may have enriched the transcript and disadvantaged handwritten notes when scored against an encounter-derived reference standard. Because the encounter audio/transcript both informed AI note generation and underpinned scoring, the comparison is not fully symmetric and handwritten notes may also draw on tacit clinical knowledge not captured in audio. In addition, the study was not powered for rare safety outcomes. These findings should therefore be interpreted as exploratory and validated in larger, multi-site studies.

This is the first reported real-world evidence that, within an ambient-scribe-enabled workflow, an LLM-based AI scribe can generate clinical notes that are more complete and less prone to serious clinical errors than HW notes in a LMIC hospital environment. By introducing a reproducible, domain-aligned evaluation framework, this study provides the groundwork for future research and supports the potential for AI scribes to improve documentation quality, accuracy, and efficiency, while emphasising the essential role of human oversight. In LMIC settings, ambient AI scribes could complement existing documentation workflows and may represent a core component of a broader pathway toward scalable digital health infrastructure.

4. Methods

4.1 Study design

This was a prospective, real-world evaluation conducted in the Division of Neurosurgery at Groote Schuur Hospital (Cape Town, South Africa). The objective of this study was to examine the properties of AI-generated notes compared with standard HW documentation, focusing on quality, safety, error profiles, and completeness. Consecutive encounters occurred across the hospital and included the trauma and emergency unit, the neurosurgical outpatient department, wards and operating rooms.

The ambient AI scribe was developed by GraiLabs. GraiLabs provided technical support limited to configuration and deployment. Clinical encounter selection, reference standard creation, outcome scoring, and all statistical analyses were performed by the academic study team.

4.2 Participants and ethics

Eligible participants were older than 17, fluent in English, able to provide informed consent, and willing to have the encounter recorded. The study was approved by the University of Cape Town Human Research Ethics Committee (HREC 241/2025) and was conducted in accordance with the ethical principles of the Declaration of Helsinki. Official documentation for the patient remained the HW clinical note or operation note, stored in the physical folder.

4.3 AI scribe system and security

Recordings were captured on a single password-protected device. Each clinician used a unique login and entered only the patient folder number prior to recording; names and other identifiers were not entered. The audio uploaded automatically to a secure Microsoft Azure cloud database in South Africa with encryption in transit and at rest, with restricted access, and audit logs. After transcription of the audio the system accessed the OpenAI GPT-5 model via a private application programming interface (API) to generate SOAP notes without any additional clinical context or templates, beyond the prompt and the transcript. Model configurations were fixed and did not change throughout the study period. Clinicians were blinded to the transcripts and AI notes throughout and could not amend or alter the audio, nor could they replay it. The interface displayed only a confirmation of successful upload. The Protection of Personal Information Act (POPIA) requirements for consent, security safeguards, and data residency were followed.[16]

4.4 Prompt optimisation

A first encounter, prospectively excluded from analysis to avoid over-fitting, was used to iteratively optimise the prompt and context, so that output matched neurosurgical documentation conventions, with prioritisation of safety guardrails, factual accuracy, adherence to local language conventions and first-person narration consistent with clinician voice. The consultation transcript was the only permissible source of clinical facts; the system was instructed not to fabricate, infer, assume, or embellish information. Limited inference of medical terminology was allowed when the transcript contained unambiguous lay descriptions or descriptive phrasing, provided that meaning was preserved. If a clinically relevant element was not mentioned, the note was required to state this.

4.5 Clinician conduct

Clinicians were asked to conduct encounters as usual. To support capture of clinically relevant content by the ambient scribe, clinicians were encouraged to verbalise key examination findings, imaging interpretations, and management decisions when these would otherwise be documented in writing. This instruction was intended to approximate a real-world ‘ambient scribe-enabled’ workflow in which clinicians may externalise elements of clinical reasoning that are often implicit or written.

4.6 Handwritten note workflow

Handwritten notes were produced using the team’s usual conventions and time constraints. Immediately after each encounter, the clinician wrote their note and recorded the time taken. Consultation duration was derived from audio-recording timestamps. HW notes were later assessed for legibility and for medico-legal completeness, specifically if the following were documented: presence of location, date, time, a legible clinician name, and signature. The scoring process is described below.

4.7 Note structure and scoring

Given the paucity of validated metrics for low-resource settings, the scoring rubric (Table 2) was developed a priori by the raters to provide a pragmatic, structured assessment framework based on the established SOAP documentation format used locally for the clinical encounters and operative notes evaluated in this study. Domains and items were informed by routine clinical documentation requirements and refined through consensus before scoring.[17]

For consultation notes, raters scored five domains (Subjective, Objective, Assessment, Plan, and Overall organisation) on a Likert scale from 0 to 5, yielding a total score from 0 to 25. The Overall organisation domain evaluated structure, flow, and internal consistency. For operative notes, the Subjective domain did not apply and the total score therefore ranged from 0 to 20. Secondary outcomes included efficiency (time to generate AI notes vs HW notes), content density (word counts of AI summaries and HW notes), HW note legibility, accuracy, completeness of ICD-10 secondary diagnostic coding, and a severity-graded error taxonomy.

<u>Score</u>	<u>Anchor label*</u>	<u>Operational definition</u>
0	Absent / dangerously incorrect	Content absent or contains dangerously incorrect information; critical elements missing such that safe care is compromised.
1	Severe deficiencies	Major omissions and/or marked disorganisation that could impair safe clinical care.
2	Partial content with notable gaps	Partial documentation with notable gaps; meaning, completeness, or chronology unclear.
3	Clinically usable with minor clarification	Generally adequate documentation that is clinically usable; requires minor clarification.
4	Good / comprehensive	Good-quality documentation that is comprehensive and internally consistent; only trivial omissions.
5	Exemplary / auditable	Complete, accurate, and well-structured documentation; includes medico-legal elements where relevant; readily auditable.

Table 2. Scoring rubric (applied identically to AI-generated and handwritten notes)

*The rubric anchors were identical for AI-generated and handwritten notes across all scored domains.

4.8 Error definitions (applied identically to AI and HW notes)

A hallucination (confabulation) was defined as the insertion of information not supported by the encounter audio recording or the medical record, presented as a clinician-attributed fact. This definition was applied to both AI-generated and handwritten notes. In handwritten notes, such content may reflect recall error or un-verbalised clinical findings, whereas in AI-generated notes it reflects model-generated

content. Clearly labelled system suggestions or prompts were not considered hallucinations if they did not misrepresent clinical intent. A meaning distortion was defined as content that had been spoken but captured in a way that materially altered its clinical meaning. A clinically significant omission was defined as the absence of a material element that would reasonably be expected and could impact care. Because the reference standard was the encounter-derived audio recording, omissions in both note types were interpreted as differences in documentation completeness relative to the verbalised recording, and may reflect appropriate clinical shorthand or local documentation conventions rather than clinical error.

To improve error granularity, each identified documentation error was scored on a five-point Likert scale ranging from 1 (minor) to 5 (catastrophic).[18] Overall error burden for each encounter was then synthesised and classified by potential clinical impact as either minor, or major.

4.9 Adjudication

Two consultant neurosurgeons independently scored all AI-generated and handwritten notes against the original audio recording, recording errors per encounter. Although AI notes were generated from automated transcripts of the encounter audio, scoring was anchored to the original audio recording, which raters used to resolve transcription errors and to classify discrepancies. Raters were instructed to assess documentation quality (rather than clinical appropriateness) when classifying errors. Raters were not blinded to note type because source-identifying features were intrinsic to the notes. Scoring was therefore performed using the predefined rubric with independent dual ratings and adjudication of high-disagreement encounters.

Inter-rater disagreement was quantified across SOAP domains, error counts, and ICD-10 coding quality. The ten encounters with the greatest disagreement underwent adjudication by a third senior neurosurgeon. For these encounters, a consensus score replaced the original ratings; all others retained their independent scores.

To ensure consistent error classification, all three neurosurgeons jointly reviewed all encounters to finalise error type and severity. Error rates for hallucinations, omissions, and distortions were calculated using these panel-reviewed determinations only. SOAP scores were not altered during this process.

4.10 Statistical analysis

Error counts for AI and HW notes were modelled separately using Bayesian Poisson rate models with weakly informative Gamma (0.5, 0.5) priors for per-case error rates. Comparisons were summarised using rate ratios of HW to AI notes (HW/AI), reported as posterior medians with 95% credible intervals and posterior probabilities that HW notes exhibited higher error rates than AI notes.

Error severity distributions (none, minor, major) were analysed using Dirichlet-Multinomial models with Dirichlet (0.5, 0.5, 0.5) priors that yield stable inference under sparse counts. We report category-level HW/AI risk ratios with 95% credible intervals. Win-rate was analysed using a Bayesian Binomial model with a Beta (0.5, 0.5) prior. Ties were handled by counting ties as half-wins (pre-specified), and posterior summaries were reported as the posterior median, 95% credible interval, and the posterior probability that the win probability exceeded 0.5. Posterior inference used Monte Carlo simulation with at least 200,000 sample draws. Inter-rater reliability across all encounters was quantified using weighted Cohen's κ for ordinal domain scores and the intraclass correlation coefficient (ICC; two-way random, absolute agreement) for continuous composite scores and error counts.

Declarations

5. Data Availability

The data generated and analysed during this study are available from the corresponding author on reasonable request. All data supporting the findings are included in the manuscript and accompanying supplementary information.

6. Code availability

No new large language model code was developed specifically for this study. The AI scribe evaluated in this work was developed by Grailabs and is not publicly available. The prompts used in the study are described in the Methods section. As detailed in the Methods, OpenAI's GPT-5 was used as the large language model for all experiments.

8. Acknowledgements

The authors thank the patients and colleagues in the Division of Neurosurgery for their participation in this study. We acknowledge Dr Gareth Obery, Dr Zameer Brey, Dr Hao Hu and Dr Scott Mahoney for valuable advice, and Stewart Truswell for support and insightful contributions. No funding has been received for this study.

9. Author contributions

B.D.J. contributed to study conception, led data collection, participated in scoring and analysis, and drafted the manuscript. J.M.N.E. contributed to study conception, data collection, participated in scoring and analysis and manuscript revision. J.F. contributed to methodology, analysis, and manuscript revision. L.C. provided technical support for system configuration and deployment. B.B. contributed to study conception, to study methodology, statistical analysis, supervision, and manuscript revision. G.F. contributed to study conception, participated in scoring and analysis, supervised the study and revised the manuscript. B.B. and G.F. are shared senior authors.

10. Competing Interest Disclosure

Linda Camara is employed by GraiLabs (Head of Product Development) and developed the ambient AI scribe system evaluated in this study, providing technical support for system configuration and deployment only and was not involved in study design, outcome assessment, statistical analysis, interpretation, or manuscript drafting. All other authors declare no competing interests.

References

1. Adams, L., et al., Artificial intelligence in health, health care, and biomedical science: an AI code of conduct principles and commitments discussion draft. *NAM perspectives*, 2024. 2024: p. 10.31478/202403a.
2. Tangsrivimol, J.A., et al., Artificial Intelligence in Neurosurgery: A State-of-the-Art Review from Past to Future. *Diagnostics (Basel)*, 2023. 13(14).
3. Patil, A., et al., Large language models in neurosurgery: a systematic review and meta-analysis. *Acta Neurochir (Wien)*, 2024. 166(1): p. 475.
4. Topaz, M., L.M. Peltonen, and Z. Zhang, Beyond human ears: navigating the uncharted risks of AI scribes in clinical practice. *NPJ Digit Med*, 2025. 8(1): p. 569.
5. Sasseville, M., et al., The Impact of AI Scribes on Streamlining Clinical Documentation: A Systematic Review. *Healthcare (Basel)*, 2025. 13(12).
6. Kanaparthi, N.S., et al., Real-World Evidence Synthesis of Digital Scribes Using Ambient Listening and Generative Artificial Intelligence for Clinician Documentation Workflows: Rapid Review. *JMIR AI*, 2025. 4: p. e76743.
7. Lee, C., S. Britto, and K. Diwan, Evaluating the Impact of Artificial Intelligence (AI) on Clinical Documentation Efficiency and Accuracy Across Clinical Settings: A Scoping Review. *Cureus*, 2024. 16(11): p. e73994.
8. Wang, H., et al., An evaluation framework for ambient digital scribing tools in clinical applications. *NPJ Digit Med*, 2025. 8(1): p. 358.
9. Olson, K.D., et al., Use of Ambient AI Scribes to Reduce Administrative Burden and Professional Burnout. *JAMA Netw Open*, 2025. 8(10): p. e2534976.
10. Ma, S.P., et al., Ambient artificial intelligence scribes: utilization and impact on documentation time. *J Am Med Inform Assoc*, 2025. 32(2): p. 381–385.
11. Pearlman, K., et al., Use of an AI Scribe and Electronic Health Record Efficiency. *JAMA Netw Open*, 2025. 8(10): p. e2537000.
12. Duggan, M.J., et al., Clinician Experiences With Ambient Scribe Technology to Assist With Documentation Burden and Efficiency. *JAMA Netw Open*, 2025. 8(2): p. e2460637.
13. Kim, E., V.X. Liu, and K. Singh, AI Scribes Are Not Productivity Tools (Yet). 2025, Massachusetts Medical Society. p. Ale2501051.
14. Lukac, P.J., et al., A Randomized-Clinical Trial of Two Ambient Artificial Intelligence Scribes: Measuring Documentation Efficiency and Physician Burnout. *medRxiv*, 2025.

15. Afshar, M., et al., A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. *NEJM AI*, 2025. 2(12): p. Aloa2500945.
16. South Africa. Protection of Personal Information Act 4 of 2013. Government Gazette., 26 November 2013.
17. Stetson, P.D., et al., Assessing Electronic Note Quality Using the Physician Documentation Quality Instrument (PDQI-9). *Appl Clin Inform*, 2012. 3(2): p. 164–174.
18. Asgari, E., et al., A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *NPJ Digit Med*, 2025. 8(1): p. 274.
19. Mess, S.A., A.J. Mackey, and D.E. Yarowsky, Artificial Intelligence Scribe and Large Language Model Technology in Healthcare Documentation: Advantages, Limitations, and Recommendations. *Plast Reconstr Surg Glob Open*, 2025. 13(1): p. e6450.
20. Bracken, A., et al., Artificial Intelligence (AI) - Powered Documentation Systems in Healthcare: A Systematic Review. *J Med Syst*, 2025. 49(1): p. 28.
21. Lodge, W., et al., Assessing completeness of patient medical records of surgical and obstetric patients in Northern Tanzania. *Glob Health Action*. 2020, Taylor & Francis.
22. Lindo, J., et al., An audit of nursing documentation at three public hospitals in Jamaica. *Journal of Nursing Scholarship*, 2016. 48(5): p. 499–507.
23. Ravikumar, K., AI scribes and digital colonialism: learning from the past to regulate the future. *BMJ*, 2025. 390: p. r2005.

Figures

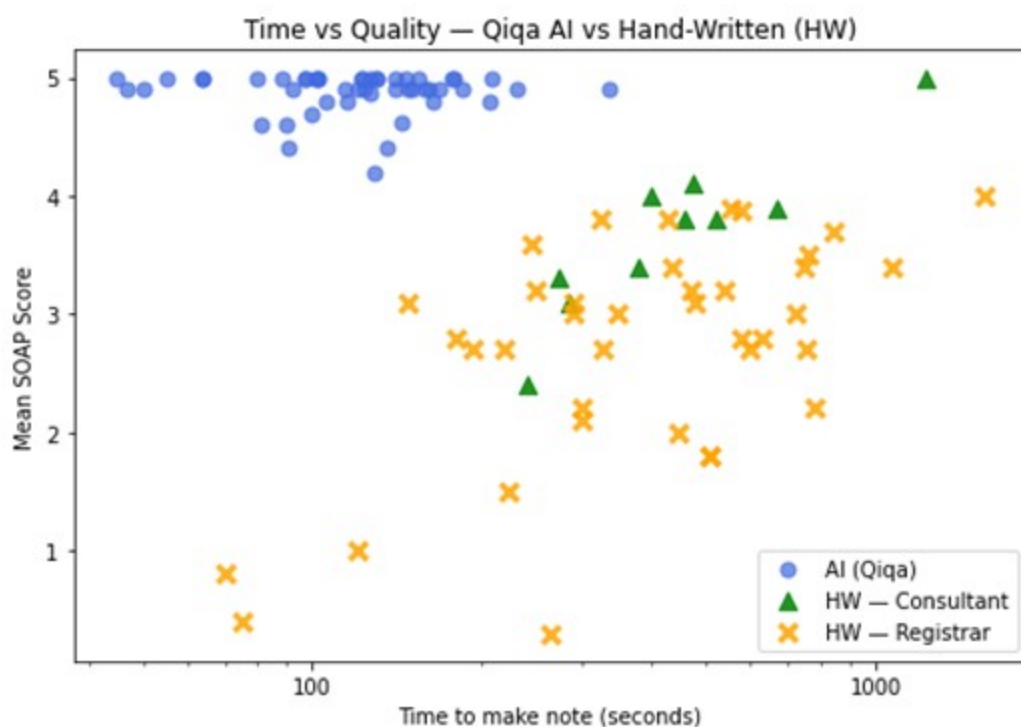


Figure 1

Graph of individual overall rubric SOAP scores vs time to make notes

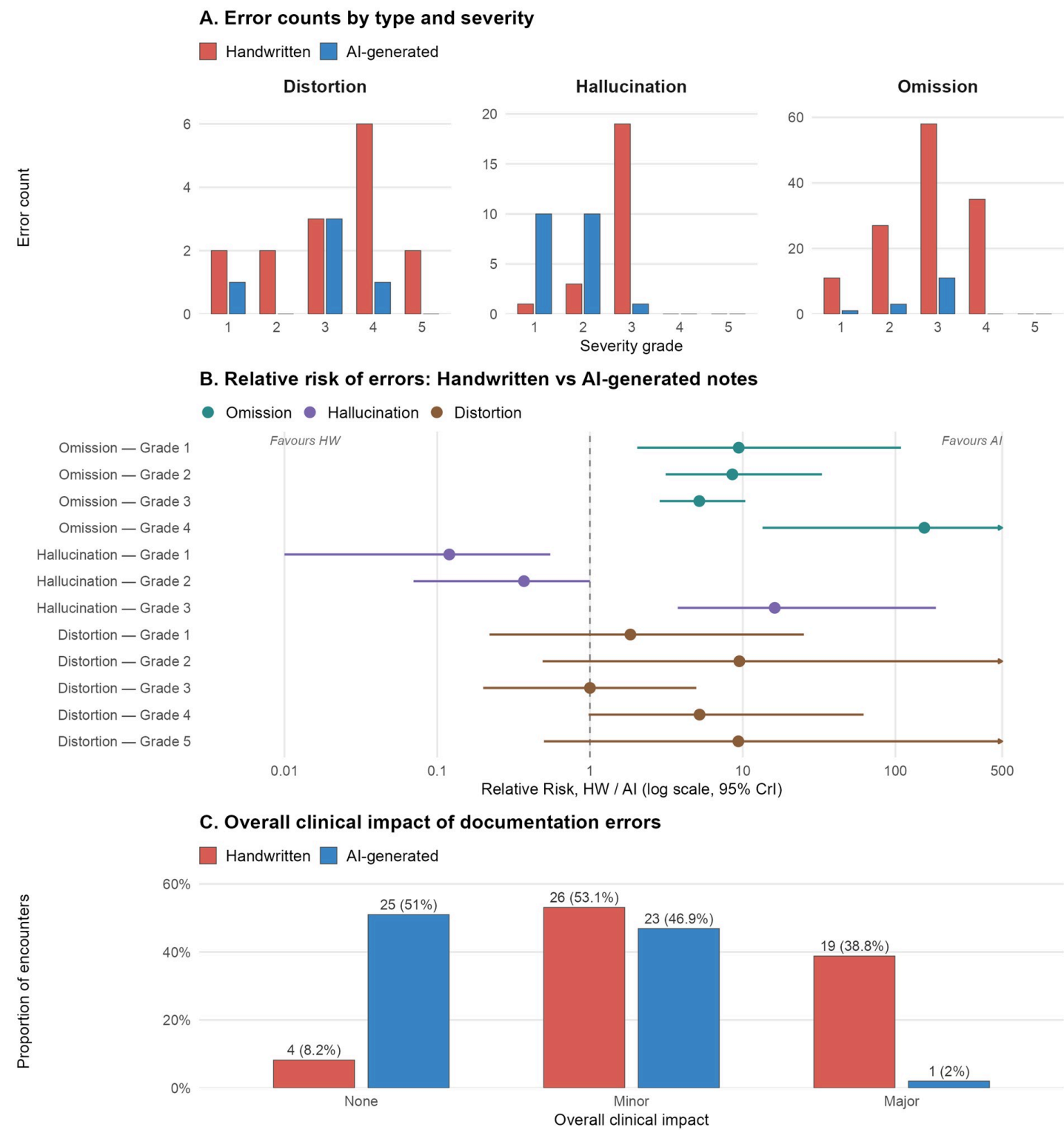
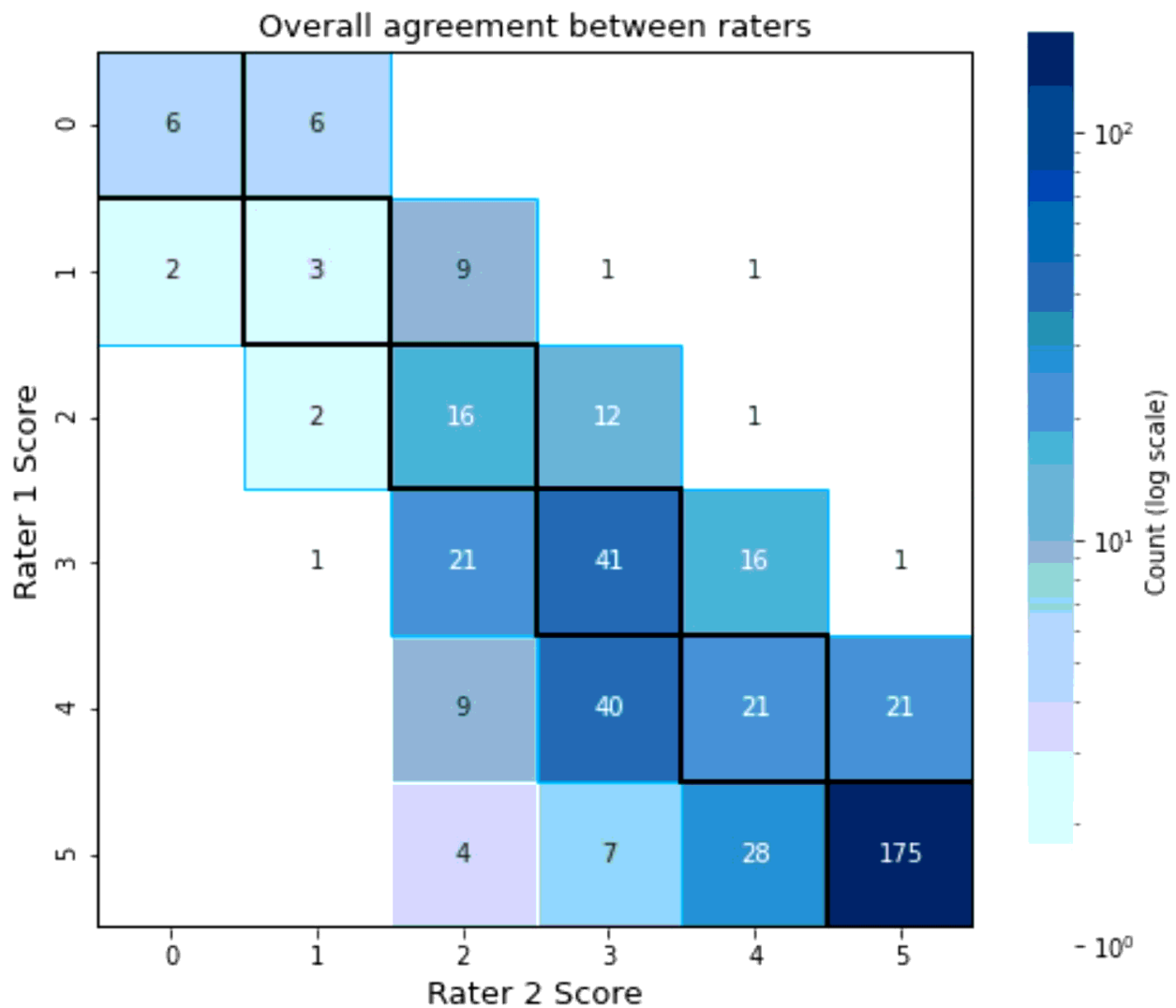


Figure 2

Error profiles and clinical impact of AI-generated versus handwritten clinical documentation. (A) Raw error counts by type (distortion, hallucination, omission) and severity grade (1 = minor, 5 = catastrophic) for AI-generated (blue) and handwritten (red) notes across 49 encounters. Omissions were the dominant error mode in handwritten documentation; AI hallucinations were predominantly low-grade (grades 1–2), whereas handwritten hallucinations clustered at grade 3. (B) Bayesian relative risk estimates (HW/AI) with 95% credible intervals on a log scale. Values greater than 1 indicate higher error rates in handwritten notes. Arrows denote truncated credible intervals. (C) Overall clinical impact classification per encounter. Major documentation errors occurred in 38.8% of handwritten notes compared with 2.0% of AI-generated notes (relative risk 16.1; 95% CrI 4.0–175.4; posterior probability 1.0). Error definitions and severity grading are described in Section 2.8. CrI = credible interval; HW = handwritten.



Quadratic Cohen kappa Exact Agreement Within-1 Agreement

0.814

0.59

0.944

Figure 3

Cross-tabulation of raw counts of rater scores showing good concordance of scores between the two raters. In 94% of cases the rater scores differed by one point or less.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterial.docx](#)