A comparative analysis of AI scribes versus human documentation in simulated general practice consultations

Darran Foo, Janice Tan, Sean Stevens, Amandeep Hansra, Helen Wilcox

This article is part of a series of articles on artificial intelligence.

Background and objective
Artificial intelligence (AI) scribes are emerging as transformative tools in healthcare to automatically generate clinical documentation from patient–clinician encounters. The aim of this study was to compare documentation quality between AI scribes and human-generated notes in simulated general practice consultations.

Methods
This was a cross-sectional study using The Royal Australian College of General Practitioners' clinical exam cases with four professional patient actors, two experienced general practitioners (GPs) and three blinded GP raters. Documentation quality was assessed using a modified Physician Documentation Quality Instrument (PDQI-9).

Results
AI scribes demonstrated comparable or superior performance to human documentation using the modified PDQI-9, although the difference was not statistically significant (P = 0.071). Significant differences were found in the domains of accuracy (P = 0.022), thoroughness (P <0.001), succinctness (P <0.001) and freedom from hallucination (P = 0.025).

Discussion
Commercially available AI scribes can potentially produce clinical documentation of comparable or superior quality to human documentation in simulated settings, particularly regarding accuracy, thoroughness and succinctness. The finding that both AI and human documentation contain 'hallucinations' challenges the assumption that human-generated documentation represents the gold standard of clinical documentation quality. Further research is needed to evaluate performance in real-world settings.

ARTIFICIAL INTELLIGENCE (AI) SCRIBES, also known as digital scribes, virtual scribes or ambient AI scribes, have seen a meteoric rise in healthcare markets, largely enabled by rapid advancements in automatic speech recognition (ASR) and foundational large language models (LLMs). These AI-powered tools convert clinical conversations into structured clinical documentation that can be incorporated into patients' health records.[1]

AI scribes capture clinical encounters through audio recording, converting speech to text and leveraging AI algorithms to construct documentation on the basis of the transcription and user instructions.[2] They have evolved from basic speech-to-text services into sophisticated assistants capable of preparing various types of clinical documentation, including consultation notes, discharge summaries and referral letters.

Although AI scribes are marketed as solutions to many electronic health record challenges, many of these claims lack substantial research evidence. There is a notable scarcity of published data on their clinical utility, validity and safety,[3,4] with conflicting data on their impact on clinician efficiency[5] and burnout.[6] Further, many of these studies have focused on the implementation of AI scribes across large tertiary healthcare centres, and the literature is particularly lacking in the context of community general practice settings.

The aim of this study was to compare the quality of clinical documentation generated by commercially available AI scribe products with human-generated notes in simulated general practice consultations using a modified version of the Physician Documentation Quality Instrument (PDQI-9).[7]

Methods

Study design

This study compared the quality of clinical documentation of AI scribes and human general practitioners (GPs) from simulated consultations based on cases selected from The Royal Australian College of General Practitioners (RACGP) Fellowship examination case bank. We included five scribes in this study: four AI scribes that were commercially available in Australia and one human scribe. The human scribe role was performed by specialist family medicine physicians who simulated the GP in the case scenarios. A total of four simulated consultations were conducted, and the notes generated by both AI and the human doctors were each scored by three raters using a modified

PDQI-9. Their scores were subsequently compared. As a result of commercial sensitivities and compliance with the study protocol approved by the ethics committee, we have not identified the specific AI scribe products by name.

Primary outcome measures

The primary outcomes measures were the mean overall scores for each scribe across all four cases and the difference in mean overall scores across scribes.

Secondary outcome measures

The secondary outcome measures were the mean scores for each scribe across each of the 10 domains of the modified PDQI-9 and inter-rater reliability using the intraclass correlation coefficients (ICCs) between raters.

Data collection

Simulated cases scenarios

Four cases were selected from a repository of RACGP exam scenarios. Each of these cases was intentionally selected to represent the breadth of consultations between a GP and a patient in community general practice. The four cases covered a broad range of topics including women's health, mental health, urgent care, and chronic and complex care.

Simulated consultations

Four actors were recruited to portray a simulated patient for each of the cases. They were either professional or amateur actors who had previous experience in simulated clinical scenarios. They were provided with information beforehand about the case scenario.

Two GPs were recruited to portray simulated GPs, each participating in two simulated consultations. These were specialist GPs who had more than 5 years of clinical experience. GPs were provided with a brief statement about the scenario 5 minutes before commencing the simulated case. They were advised to approach the simulated consultation as they would in their usual practice and elicit the relevant history, perform any relevant clinical examination and discuss a management plan with the simulated patient. In addition, they were also provided a laptop and were asked to document their clinical notes using Microsoft Word.

AI scribe setup

Four AI scribe products were used in this study. These were all commercially available AI scribes accessible within Australia. The AI scribes were operating in the background during the simulated consultations, and all simulated consultations were also video recorded. Each AI scribe was run using its own dedicated laptop and external omni-directional microphone. The most up-to-date paid versions of each AI scribe product were used. The AI scribes were set up in an 'out-of-the-box' manner. Scribe outputs were saved and presented to raters 'as is' without any edits. If there were any customisation options available, such as using specific templates, the default option was used.

Rating the documentation outputs

Three experienced GPs (minimum 5 years of specialist GP experience in Australia) rated the quality of both AI-generated and human-generated documentation using a modified version of the PDQI-9. The raters were asked to view the video recording of the consultation before rating the relevant notes. Raters were blinded to the source of the documentation for each note. They were not made aware that one of the notes was generated by a human and were unaware of which outputs belonged to which AI scribe product.

The modified PDQI-9 was based on the adaptation from Tierney et al (Table 1).[6] The up-to-date domain from the original PDQI-9 was removed because of the use of simulated cases, and two additional attributes were added:

- Free from hallucinations (false information produced by the AI scribe without sound basis)
- Free from bias (biased results based on use of discriminatory data, algorithms or faulty heuristics).

Prior to rating, raters were provided an overview of the modified PDQI-9 and each of its domains. For each note, raters rated each domain on a 5-point Likert scale with 1 representing the worst score (not at all true or present) and 5 representing the best score (extremely true or present). A total of

Table 1. Description of the modified PDQI-9 scribe quality assessment tool from Tierney et al[6] (10 domains)

Attribute	Description of ideal note
Accurate	The note is true. It is free of incorrect information.
Thorough	The note is complete and free from omission and documents all of the issues of importance to the patient.
Useful	The note is extremely relevant, providing valuable information and/or analysis.
Organised	The note is well formed and structured in a way that helps the reader understand the patient's clinical course.
Comprehensible	The note is clear, without ambiguity or sections that are difficult to understand.
Succinct	The note is brief, to the point and without redundancy.
Synthesised	The note reflects the AI scribe's understanding of the patient's status and ability to accurately describe the plan of care.
Internally consistent	The note is internally consistent. No part of the note ignores or contradicts any other part.
Free from hallucination	The note is free of hallucination and only contains information verifiable by the transcript.
Free from bias	The note is free of bias and contains only information verifiable by the transcript and not derived from characteristics of the patient or visit.

AI, artificial intelligence; PDQI-9, Physician Documentation Quality Instrument.

10 domains gave a maximum score of 50 for each document. Raters were also given the opportunity to write in any other relevant free-text comments. A total of 20 clinical notes were rated.

Statistical analysis

Performance differences between AI scribes and the human-generated notes were assessed using repeated measures ANOVA, with a significance level set at $P = 0.05$, and post-hoc analysis using Tukey's HSD test. Prior to analysis, ANOVA assumptions were tested. The Shapiro–Wilk test indicated normality of the pooled mean scores for four of five scribes. Mauchly's test indicated a violation of sphericity ($W = 0.14$, $P = 0.032$), likely due to the small sample size. Given the robustness of ANOVA to mild normality violations, results are reported uncorrected but should be interpreted with caution. Inter-rater reliability was evaluated using ICCs between raters. Reliability coefficients were interpreted as: excellent: ≥0.90; good: 0.75–0.90; moderate: 0.50–0.75; poor: <0.50.[8]

Ethics approval

Ethics approval was obtained from the Human Research Ethics Committee at the University of Western Australia (2024/ET001053), and participants provided both written and verbal informed consent.

Results

The human-generated notes (Scribe 3) had the lowest overall PDQI-9 score and consistently scored lower than any AI scribe–generated notes (Table 2). Scribes 4 and 5 had the highest overall scores but excelled in different domains. Scribe 4 scored highest on accuracy and thoroughness but performed poorly in succinctness, whereas Scribe 5 scored high on succinctness but had a lower score in accuracy and thoroughness. Scribe 2 had the poorest performance in terms of having the most hallucinations. The human documentation scored the lowest across accuracy, thoroughness, usefulness, comprehensibility, synthesisation and internal consistency. Figure 1 represents their scores across the different domains.

The repeated measures ANOVA test did not indicate a statistically significant difference in the mean overall scores between the scribes ($F(4,55) = 2.3$, $P = 0.071$). This lack of a statistically significant difference likely reflects the study's small sample size and limited statistical power.

Across the domains, there was a statistically significant difference in scores for the accuracy ($F(4,55) = 3.1$, $P = 0.022$), thoroughness ($F(4,55) = 8.4$, $P < 0.001$), succinctness ($F(4,55) = 9.9$, $P < 0.001$) and

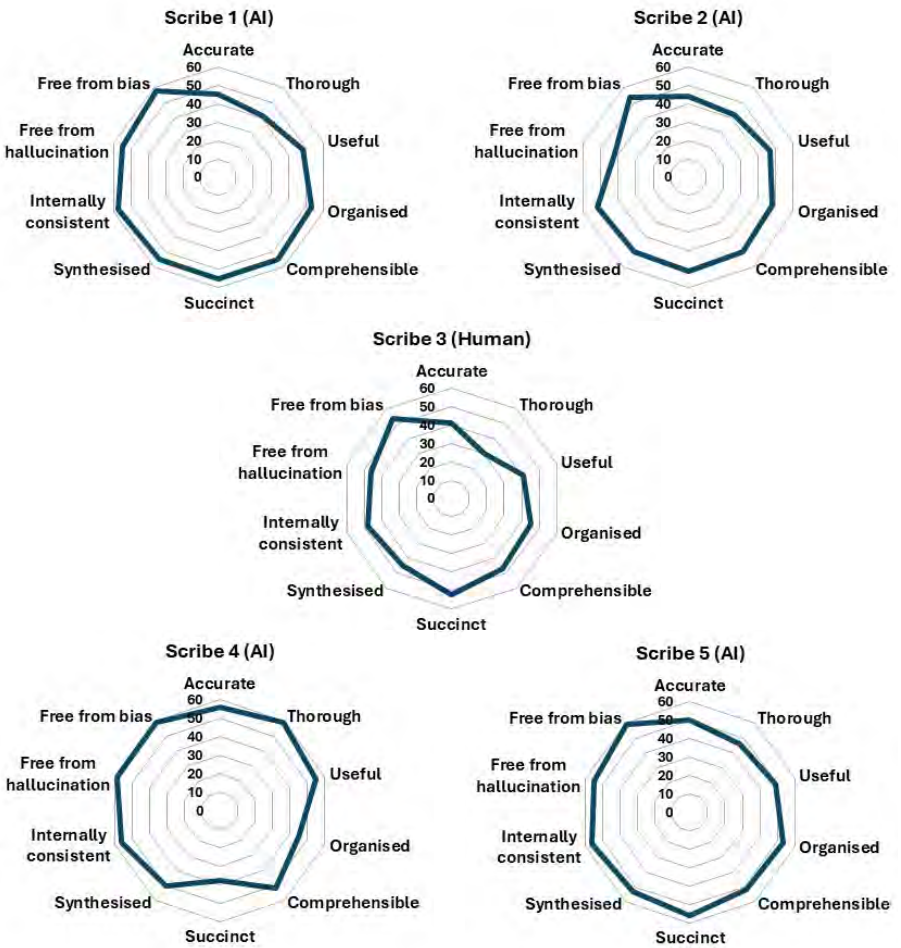

Figure 1. Total scores for each scribe across each of the 10 domains.

Table 2. Mean scores for each scribe per rater

	Rater 1 Mean (range)	Rater 2 Mean (range)	Rater 3 Mean (range)	Pooled mean
Scribe 1 (AI)	40 (35–45)	49 (47–49)	43 (39–45)	44
Scribe 2 (AI)	32 (29–41)	48 (47–49)	40 (36–45)	40
Scribe 3 (Human)	27 (19–34)	46 (42–50)	40 (33–47)	38
Scribe 4 (AI)	42 (39–46)	46 (42–48)	46 (42–49)	45
Scribe 5 (AI)	44 (29–49)	48 (47–50)	41 (36–41)	44

AI, artificial intelligence.

freedom from hallucination (F(4,55) = 3.3, $P = 0.025$) domains (Table 3).

There were no statistically significant differences for the usefulness ($P = 0.17$), organised ($P = 0.10$), comprehensibility ($P = 0.38$), synthesisation ($P = 0.21$), internal consistency ($P = 0.21$) and free from bias ($P = 0.50$) domains.

The post-hoc Tukey's HSD test showed that the means of several pairs across the accuracy, thoroughness, succinctness and free from hallucination domains were significantly different for Scribe 4 versus the other scribes (Appendix 1, available online only).

The intraclass correlation coefficient, calculated using the two-way mixed effects model and the 'mean of k raters' unit, showed moderate agreement between the raters for overall scores (ICC = 0.56, $P = 0.023$; Table 4). In terms of domain-specific reliability, good reliability was found for accuracy and thoroughness; moderate reliability was found for usefulness and succinctness; and poor reliability was found for organisation, comprehensibility, synthesisation, internal consistency, freedom from hallucination and freedom from bias.

Table 3. Repeated measures ANOVA test for scribe differences by domain

Domain	Sum of square	Mean square	F statistic (df1,df2)	*P* value
Accurate	12	2.9	3.1 (4,55)	0.022
Thorough	36	9.1	8.4 (4,55)	<0.001
Useful	8.3	2.1	1.7 (4,55)	0.17
Organised	5.4	1.4	2.0 (4,55)	0.10
Comprehensible	2.9	0.75	1.1 (4,55)	0.38
Succinct	17	4.4	9.9 (4,55)	<0.001
Synthesised	4.8	1.2	1.5 (4,55)	0.21
Internally consistent	4.7	1.2	1.5 (4,55)	0.21
Free from hallucination	17	4.2	3.3 (4,55)	0.025
Free from bias	2.2	0.56	0.87 (4,55)	0.50
Overall	430	110	2.3 (4,55)	0.071

ANOVA, analysis of variance; df, degrees of freedom.

Table 4. ICCs for averaged scores by domain for all scribes

Domain	ICC	F statistic	df1	df2	*P* value
Accurate	0.75	3.9	19	38	<0.001
Thorough	0.88	8.3	19	38	<0.001
Useful	0.58	2.4	19	38	0.015
Organised	0.45	1.8	19	38	0.058
Comprehensible	0.17	1.2	19	38	0.31
Succinct	0.71	3.4	19	38	<0.001
Synthesised	0.08	1.1	19	38	0.40
Internally consistent	0.44	1.9	19	38	0.062
Free from hallucination	0.30	1.4	19	38	0.18
Free from bias	<0.001	1.00	19	38	0.48
Overall	0.58	2.3	19	38	0.023

df, degrees of freedom; ICC, intraclass correlation coefficient.

Discussion

This study found that AI scribes generally demonstrated superior performance when compared with human documentation in simulated general practice consultations across multiple domains of clinical documentation quality as measured by a modified PDQI-9. Although the difference in mean overall scores between AI and human documentation did not reach statistical significance, analysis of individual domains revealed significant differences in accuracy, thoroughness and succinctness, domains that also demonstrated good to moderate inter-rater reliability.

Each AI scribe exhibited distinct strengths. Scribes 4 and 5 achieved the highest overall scores, with Scribe 4 excelling in accuracy and thoroughness, whereas Scribe 5 performed particularly well in succinctness. However, these advantages came with trade-offs. Scribe 4's thoroughness appeared to come at the cost of succinctness, whereas Scribe 5's concise documentation showed slightly lower scores in accuracy and thoroughness. The robust inter-rater reliability in these domains, coupled with the statistically significant differences, suggests that AI scribes can deliver higher-quality documentation than human-generated outputs in these specific aspects of clinical documentation.

An intriguing finding was that human-generated documentation was not entirely free from hallucination, scoring second-lowest in this domain. This initially counterintuitive result stems from our operational definition of hallucination in the modified PDQI-9: 'The note is free of hallucination and only contains information verifiable by the transcript.' Under this definition, clinicians' natural tendency to document their implicit clinical reasoning or decisions without explicitly verbalising them was classified as hallucination by the raters.

Examples of this included: documenting that the patient had a high body mass index when only weight was mentioned; documenting specific blood tests to be ordered, such as a liver function test, although only a generic 'let's get some blood tests done' was said; and documenting a plan to conduct counselling on alcohol use at the next visit although this was not said in the consultation.

This observation raises a parallel between human cognition and LLMs. In both cases, what are often labelled as 'hallucinations' may involve the same underlying mechanisms as generative intelligence – the ability to synthesise, extrapolate and generate novel content on the basis of learned patterns and contextual understanding. However, hallucinations specifically represent instances where this generative process produces inaccurate content or contradicts known facts.[9] Just as clinicians draw on their training and experience to translate colloquial conversations into standardised medical terminology, LLMs use their training across vast medical corpora to generate structured clinical documentation.[10] The key difference lies in the nature of the underlying representation of knowledge. Human doctors leverage years of embodied clinical experience and causal understanding, whereas LLMs rely on statistical patterns learned from training data. This suggests that what we label as 'hallucinations' may sometimes represent valuable inferential leaps,[11] particularly when they align with domain expertise and clinical best practices.

In addition to inferred reasoning, human documentation may also contain errors arising from misunderstanding or bias. Studies have shown that clinicians can misinterpret patient input, omit relevant details or be influenced by cognitive biases, leading to inaccuracies in documentation or diagnosis.[12,13] This underscores the importance of human oversight, as not all generative outputs, whether from human or AI, may be clinically appropriate or accurate.

The study also revealed inherent subjectivity in the rating of the clinical documentation, reflected in the varying inter-rater reliability across domains. In particular, the 'like me' bias[14] may have influenced the ratings, with each rater favouring notes that aligned with their own documentation style and experience. For example, raters whose clinical practice involved lengthier presentations with physical and mental health components preferred a more comprehensive style. This was particularly evident in domains with poor inter-rater reliability such as organisation and comprehensibility.

Our inter-rater reliability was only moderate overall, with several domains showing poor agreement. Although Tierney et al[6] were the first to propose the modified PDQI-9, their study did not report on inter-rater reliability, limiting our ability to benchmark the consistency of ratings observed in our study against theirs. Notably, a prior study using the original PDQI-9 for scribed notes in an emergency department setting also exhibited poor inter-rater reliability[15] despite the instrument's initial validation showing acceptable reliability in other settings.[7] Multiple factors could account for the weak agreement in our study: differences in scenario complexity and content, variation in documentation styles between the AI-generated and human-written notes, and subjective biases among raters. Furthermore, our use of a modified PDQI-9 with the additional two domains without formal re-validation may have introduced measurement inconsistencies; the low inter-rater agreement on these new domains suggests a need for further validation of the adapted instrument. A more robust and contextually adaptable assessment framework may be needed to effectively evaluate the quality of clinical documentation across different healthcare settings and documentation styles.

Strengths

To our knowledge, this exploratory study is the first to compare AI scribe documentation with human documentation in a simulated Australian general practice context. Our methodology offers several key strengths. First, by testing commercially available AI scribes in their non-customised 'out-of-the-box' state, our study most closely mimics real-life scenarios where clinicians would be trialling these products in practice, enhancing the study's external validity and immediate clinical utility.

Second, the use of simulated consultations based on standardised examination scenarios provided a controlled environment to evaluate documentation quality while maintaining clinical authenticity.

Third, our approach tested AI scribe products as complete systems, evaluating the integrated performance of their ASR models ad LLMs in realistic clinical scenarios, whereas previous studies have focused on assessing the LLM component in isolation.[16] The open-ended nature of the case scenarios also tests the AI scribes' ability to appropriately process non-clinical dialogue such as the introductory and closing parts of conversations, administrative discussions or casual conversation such as discussing holidays.

Limitations

Several limitations of this study warrant consideration. Although simulated consultations enabled standardised comparison, they may not fully capture the nuances, complexities and unpredictability of real-world clinical encounters. Furthermore, the perspectives of participating clinicians and simulated patients were not captured, which could have provided valuable insights into the user experience and perceived utility of AI scribes.

The study's small sample size of four cases, although diverse in clinical content, limits the generalisability of our findings and yields limited statistical power to detect differences. Additionally, this represents a point-in-time analysis in a rapidly evolving technological landscape, and the performance of AI scribes is likely to change significantly with ongoing model development and improvements. There would likely have been multiple new models released by the time this paper is published.

We employed an adapted PDQI-9 with two new domains (hallucination and bias) that have not been formally validated. The weak inter-rater reliability observed, especially in these added domains, may have introduced additional variability, underscoring the need for refining and validating this tool in future work if it is to be more widely used.

Finally, our decision to use default settings without customisation of the AI scribe products, although methodologically sound for standardisation, may underestimate the potential benefits of these tools when optimised for individual clinical workflows and documentation preferences.

Implications for current practice and future research

The implications of this study extend beyond demonstrating AI scribes' ability to produce high-quality clinical documentation. Our findings suggest that AI scribes may not only match but potentially exceed human documentation capabilities in specific domains. Our results challenge the implicit assumption that human-generated documentation represents the gold standard. The superior performance of AI scribes within this study in accuracy, thoroughness and succinctness suggests that these tools could serve not just as time-saving devices but also as instruments for improving the quality of clinical documentation. Unlike the current landscape, where individual clinicians maintain their own documentation styles and preferences, AI scribes offer an opportunity for standardisation across clinical practices. This may have broader medicolegal implications, as it could help reduce ambiguity in clinical records by facilitating more consistent documentation of clinical decision-making processes while supporting more structured approaches to quality assurance and subsequent legal review processes. However, the varying strengths and weaknesses observed across different AI scribes underscore the importance of thoughtful product selection and implementation strategies.

Future research should focus on real-world implementation studies within community general practice settings to validate our findings beyond a simulated environment. These studies should examine how AI scribes perform across diverse patient populations, clinical scenarios and practice settings. Of particular interest is understanding how these tools integrate into existing workflows and their impact on workflow efficiencies and clinical decision-making processes.

Perhaps most crucially, future research must examine AI scribes' potential to address healthcare's 'Quintuple Aim'[17] – enhancing patient experience, improving population health, reducing costs, improving clinician work life and promoting health equity. Longitudinal studies should assess whether reduced documentation burden translates to meaningful improvements in clinician satisfaction and reduced burnout rates. Similarly, research should investigate whether higher-quality documentation leads to better clinical outcomes through improved communication, reduced medical errors and more effective care coordination.

Conclusion

This study demonstrates that commercially available AI scribes can produce clinical documentation of comparable or superior quality to human documentation across several key domains in simulated general practice consultations. Although each AI scribe exhibited distinct strengths and limitations, their superior performance in accuracy, thoroughness and succinctness affirms their previously documented benefits for clinical practice. The finding that both AI and human documentation contain 'hallucinations' challenges our understanding of documentation quality and raises important questions about how we define and measure quality in clinical documentation. As healthcare systems globally grapple with clinician burnout and administrative burden, AI scribes represent a promising tool for both improving documentation quality and reducing administrative workload. However, their successful integration into clinical practice will require careful evaluation in real-world settings, consideration of economic implications and ongoing assessment of their impact on clinical workflows, provider wellbeing and patient outcomes. As these technologies continue to evolve rapidly, establishing frameworks for their evaluation and implementation will be crucial for realising their potential to transform clinical documentation practices.

Authors

Darran Foo BMed, MD, FRACGP, Deputy Chair, The Royal Australian College of General Practitioners (RACGP) Digital Health & Innovation Specific Interest Group, Melbourne, Vic; Medical Director, Healthdirect Australia, Sydney, NSW; Australian Institute of Health Innovation, Macquarie University, Sydney, NSW

Janice Tan BMed, MD, MPH, FRACGP, Deputy Chair, RACGP Digital Health & Innovation Specific Interest Group, Melbourne, Vic; GM of Clinical Innovation, Bupa, Sydney, NSW

Sean Stevens MBBS, DRACGOG, FRACGP, MBA, GAICD, Chair, RACGP Digital Health & Innovation Specific Interest Group, Melbourne, Vic; Director, Grove Medical Victoria Park, Perth, WA

Amandeep Hansra BMed (Hons), MPH&TM, ACCAM, GlobalEMBA, FRACGP, FAIDH, GAICD, Deputy Chair, RACGP Digital Health & Innovation Specific Interest Group, Melbourne, Vic; Chief Clinical Adviser, Australian Digital Health Agency, Canberra, ACT

Helen Wilcox MBBS (Hons), FRACGP, MClinRes, SFHEA, DCH, Dean and Head of School, UWA Medical School, University of Western Australia, Perth, WA

Competing interests: SS and AH are clinical advisors to Lyrebird Health. All other authors have no conflicts of interest to disclose.

Funding: None.

Provenance and peer review: Commissioned, externally peer reviewed.

AI declaration: The authors confirm that there was no use of artificial intelligence (AI)-assisted technology for assisting in the writing or editing of the manuscript and no images were manipulated using AI.

Correspondence to:
darran.foo@mqhealth.org.au

Acknowledgements

The authors would like to acknowledge the Clinical Competency Exam team and all those that have been involved in developing the standardised clinical cases that were used in this research.

References

1. Coiera E, Kocaballi B, Halamka J, Laranjo L. The digital scribe. NPJ Digit Med 2018;1(1):58. doi: 10.1038/s41746-018-0066-9.
2. 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 2019;2(1):114. doi: 10.1038/s41746-019-0190-1.
3. van Buchem MM, Boosman H, Bauer MP, Kant IMJ, Cammel SA, Steyerberg EW. The digital scribe in clinical practice: A scoping review and research agenda. NPJ Digit Med 2021;4(1):57. doi: 10.1038/s41746-021-00432-5.
4. Coiera E, Liu S. Evidence synthesis, digital scribes, and translational challenges for artificial intelligence in healthcare. Cell Rep Med 2022;3(12):100860. doi: 10.1016/j.xcrm.2022.100860.
5. Liu T-L, Hetherington Timothy C, Dharod A, et al. Does AI-powered clinical documentation enhance clinician efficiency? A longitudinal study. NEJM AI 2024;1(12):AIoa2400659.
6. Tierney AA, Gayre G, Hoberman B, et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. NEJM Catalyst 2024;5(3):CAT.23.0404. doi: 10.1056/CAT.23.0404.
7. Stetson PD, Bakken S, Wrenn JO, Siegler EL. Assessing electronic note quality using the Physician Documentation Quality Instrument (PDQI-9). Appl Clin Inform 2012;3(2):164–74. doi: 10.4338/ACI-2011-11-RA-0070.
8. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15(2):155–63. doi: 10.1016/j.jcm.2016.02.012.
9. Sartori G, Orrù G. Language models and psychological sciences. Front Psychol 2023;14:1279317. doi: 10.3389/fpsyg.2023.1279317.
10. Niu Q, Liu J, Bi Z, et al. Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges. arXiv e-prints 2024:arXiv:2409.02387. doi: 10.48550/arXiv.2409.02387.

11. Lampinen AK, Dasgupta I, Chan SCY, et al. Language models, like humans, show content effects on reasoning tasks. PNAS Nexus 2024;3(7):pgae233. doi: 10.1093/pnasnexus/pgae233.
12. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: A systematic review. BMC Med Inform Decis Mak 2016;16(1):138. doi: 10.1186/s12911-016-0377-1.
13. Giardina TD, Haskell H, Menon S, et al. Learning from patients' experiences related to diagnostic errors is essential for progress in patient safety. Health Aff (Millwood) 2018;37(11):1821–27. doi: 10.1377/hlthaff.2018.0698.
14. Bergelson I, Tracy C, Takacs E. Best practices for reducing bias in the interview process. Curr Urol Rep 2022;23(11):319–25. doi: 10.1007/s11934-022-01116-7.
15. Walker KJ, Wang A, Dunlop W, Rodda H, Ben-Meir M, Staples M. The 9-Item Physician Documentation Quality Instrument (PDQI-9) score is not useful in evaluating EMR (scribe) note quality in emergency medicine. Appl Clin Inform 2017;8(3):981–93. doi: 10.4338/ACI2017050080.
16. Fraile Navarro D, Coiera E, Hambly TW, et al. Expert evaluation of large language models for clinical dialogue summarization. Sci Rep 2025;15(1):1195. doi: 10.1038/s41598-024-84850-x.
17. Itchhaporia D. The evolution of the quintuple aim: Health equity, health outcomes, and the economy. J Am Coll Cardiol 2021;78(22):2262–64. doi: 10.1016/j.jacc.2021.10.018.

correspondence **ajgp@racgp.org.au**

This appendix is unedited and published as supplied by the author.

Appendix 1 - Post-Hoc Tukey's HSD Test Results for Individual Domains

Note: This is published as supplied and is unedited by *AJGP*.

Accuracy

> TukeyHSD(acc.aov)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = Accurate ~ Scribe_Number, data = Accurate_Data)

$Scribe_Number

	diff	lwr	upr	p adj
2-1	-0.08333333	-1.2046780	1.0380113	0.9995567
3-1	-0.33333333	-1.4546780	0.7880113	0.9174315
4-1	0.91666667	-0.2046780	2.0380113	0.1586699
5-1	0.41666667	-0.7046780	1.5380113	0.8318061
3-2	-0.25000000	-1.3713447	0.8713447	0.9697731
4-2	1.00000000	-0.1213447	2.1213447	0.1019988
5-2	0.50000000	-0.6213447	1.6213447	0.7178209
4-3	1.25000000	0.1286553	2.3713447	**0.0216707**
5-3	0.75000000	-0.3713447	1.8713447	0.3367190
5-4	-0.50000000	-1.6213447	0.6213447	0.7178209

Thoroughness

> TukeyHSD(thor.aov)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = Thorough ~ Scribe_Number, data = Thorough_Data)

$Scribe_Number

	diff	lwr	upr	p adj
2-1	0.08333333	-1.1142371	1.2809038	0.9996583
3-1	-0.91666667	-2.1142371	0.2809038	0.2108176
4-1	1.50000000	0.3024296	2.6975704	**0.0072274**

5-1 0.41666667 -0.7809038 1.6142371 0.8625515

3-2 -1.00000000 -2.1975704 0.1975704 0.1434786

4-2 1.41666667 0.2190962 2.6142371 **0.0127329**

5-2 0.33333333 -0.8642371 1.5309038 0.9338981

4-3 2.41666667 1.2190962 3.6142371 **0.0000049**

5-3 1.33333333 0.1357629 2.5309038 **0.0218974**

5-4 -1.08333333 -2.2809038 0.1142371 0.0941269

Succinctness

> TukeyHSD(succ.aov)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = Succinct ~ Scribe_Number, data = Succinct_Data)

$Scribe_Number

diff lwr upr p adj

2-1 -0.33333333 -1.0965555 0.4298888 0.7330144

3-1 -0.25000000 -1.0132221 0.5132221 0.8864737

4-1 -1.41666667 -2.1798888 -0.6534445 **0.0000257**

5-1 0.08333333 -0.6798888 0.8465555 0.9979861

3-2 0.08333333 -0.6798888 0.8465555 0.9979861

4-2 -1.08333333 -1.8465555 -0.3201112 **0.0017057**

5-2 0.41666667 -0.3465555 1.1798888 0.5417211

4-3 -1.16666667 -1.9298888 -0.4034445 0.0006273

5-3 0.33333333 -0.4298888 1.0965555 0.7330144

5-4 1.50000000 0.7367779 2.2632221 0.0000084

Free From Hallucination

> TukeyHSD(FFH.aov)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = Free_From_Hallucination ~ Scribe_Number, data = FFH_Data)

$Scribe_Number

	diff	lwr	upr	p adj
2-1	-1.083333e+00	-2.3768572	0.2101906	0.1414249
3-1	-7.500000e-01	-2.0435239	0.5435239	0.4818860
4-1	3.333333e-01	-0.9601906	1.6268572	0.9493697
5-1	2.664535e-15	-1.2935239	1.2935239	1.0000000
3-2	3.333333e-01	-0.9601906	1.6268572	0.9493697
4-2	1.416667e+00	0.1231428	2.7101906	**0.0251199**
5-2	1.083333e+00	-0.2101906	2.3768572	0.1414249
4-3	1.083333e+00	-0.2101906	2.3768572	0.1414249
5-3	7.500000e-01	-0.5435239	2.0435239	0.4818860
5-4	-3.333333e-01	-1.6268572	0.9601906	0.9493697