Assessing the Effectiveness of an Artificial Intelligence Tool for Note-taking in a General Practice Setting

An evaluation of the Heidi AI note-taking tool

Shreya Shah, Aarya Shetye, Carol Habib
Guy's, King's and St Thomas' School of Medical Education
King's College London
London, United Kingdom
e-mail: shreya.2.shah@kcl.ac.uk, aarya.shetye@kcl.ac.uk, carol.habib@kcl.ac.uk

Abstract— **Documentation burden in general practice reduces efficiency and productivity. This is particularly important as clinicians face increasing time pressures, and inaccurate documentation can compromise patient safety and continuity of care. Large language models may assist with automated note-generation, but evidence of practical performance remains very limited. Existing studies are limited in number and often rely on simulated consultations or non-clinical settings, leaving uncertainty regarding real-world performance in primary care. This study assessed the accuracy of Heidi, an artificial intelligence medical scribe software, compared to manual note-taking by medical students in a general practice setting. Across six consultations, Heidi achieved a mean accuracy of 84%, with most errors being minor or clinically non-significant, suggesting that AI-assisted documentation may safely support clinical workflows when combined with appropriate human oversight.**

Keywords-Artificial Intelligence; General Practice; Documentation; Efficacy; Heidi; Large Language Models; Note-taking; Primary Care; Digital Health.

I. INTRODUCTION

Effective documentation is a cornerstone of healthcare, requiring both accuracy and efficiency. It is crucial for patient safety, ongoing care, and adherence to legal standards. Generally, general practitioners rely on manual note-taking either during or after a consultation. However, this can be time-intensive and may interrupt the natural flow of doctor–patient communication. Recently, artificial intelligence has become more prominent in healthcare. It is being used to support clinicians in optimising documentation without compromising the quality of patient communication. Among these tools is Heidi, a clinical documentation system powered by artificial intelligence, which is designed to improve the accuracy, consistency, and speed of consultations.

Started in 2019 [1], Heidi is a medical scribe that uses Artificial Intelligence (AI) and Natural Language Processing (NLP) to convert real-time or recorded speech into structured clinical notes. It aims to streamline administrative tasks for medical professionals by generating formatted, editable documentation aligned with clinical standards. Therefore, it gives clinicians more time to prioritise patient-centred care and minimise post-consultation administrative tasks.

There have been few studies comparing AI note-taking in comparison to manual note-taking in consultations, one of which compared ChatGPT, Heidi and manual notes in accuracy, readability and efficiency to generate dermatology consultation letters [2]. Heidi was found to be the most consistent and reliable for clinical implementation [2]. The aim of our study was to compare the efficacy of AI note-taking software with healthcare professionals' note-taking in a general practice setting in England, particularly when being utilised by medical students. It explores whether Heidi can be practically and advantageously integrated into general practice, contributing to the wider conversation about the growing role of technology in healthcare settings.

In Section 2, we outline our methods of evaluating Heidi's performance. In Section 3, we start looking at trends seen in our data which we put in a wider clinical context in Section 4, alongside beginning to predict future work and implementation.

II. METHODS

Heidi, an AI medical scribe designed to automate clinical documentation, was used for the purpose of this study. Heidi was chosen due to the ease of use, low running cost and it being one of the few AI tools being particularly developed for use in healthcare at the time. We had no prior association with nor were we approach by Heidi for the purposes of this study.

Across six consultations, AI generated and medical student generated notes were composed. Three reviewers used a four-section classification and 18-point rubric to assess the accuracy of both types of notes for each patient. Scores were also analysed qualitatively depending on the length and complexity of the consultation. Heidi produced complete notes for all six consultations. The mean accuracy score was 84% when excluding irrelevant inaccuracies, with two consultations receiving perfect scores. Most inaccuracies reflected omissions or minor contextual misunderstandings, and clinically significant inaccuracies were rare. Heidi demonstrated high accuracy and potential as an assistive documentation tool in general practice.

Before the study, test runs were conducted to assess Heidi's features and its ability to recognise different accents and multiple people talking in the same consultation. The free version of Heidi was utilised. Six patients participated in the

ISBN: 978-1-68558-362-0

study. Five in-person patients provided written informed consent to use the Heidi software, have the consultation recorded and use their consultation notes for the study. One patient who had a telephone consultation provided informed verbal consent. Heidi is in compliance with the Data Protection Act and ensures the secure management of data for both the NHS and private practice. Consultations were carried out by two medical students at a time. The primary medical student carried out the consultation while the second medical student monitored software functionality. The primary medical student took brief notes in a notebook during the consultation. They were given three minutes at the end to collate and finalise their consultation notes on a computer. This time pressure was utilised to simulate a realistic setting in a general practice. All primary medical students used this method to standardise note-taking style. Heidi was activated at the start of the consultation and stopped before the general practitioner arrived and reviewed the patient. The AI notes were transcribed and generated in real time by Heidi. The note template used on Heidi for all patients was the "General Practitioner's note". Management plans were not included or analysed in this study.

The patient consultations of this study spanned over three months, during which Heidi updated its note template and transcript formats. Variations in consultation note formats were taken into account during analysis. Three medical students independently reviewed both the AI and human-generated consultation notes. The following rubric was used: The consultation was divided into six sections:

Presenting Complaint, Past Medical History, Medication History (Including Allergies), Social History, Family History, Examination (If Applicable).

The consultation was marked under four classifications:

1. Accurate
2. Inaccurate, Will Not Make a Difference (Inaccurate WNMD)
3. Inaccurate, Will Make a Difference (Inaccurate, WMD)
4. Not Applicable (N/A).

Each section of the consultation was marked under one of the four classifications. The results were collated, and a consensus was reached after discussion.

Next, a quantitative analysis was performed. Each transcript was scored under a 3-2-1 system which was discussed and agreed upon by three reviewers. This system assigned each of the classifications points: Accurate (3 points), Inaccurate WNMD (2 points), Inaccurate WMD (1 point), N/A (3 points). A section was classified as inaccurate if there was a discrepancy in it between student notes and Heidi's notes. A section was classified as N/A if it was not covered in the consultation, and this absence was confirmed in the transcript, student notes, and Heidi's notes. N/A was assigned 3 points because these items were genuinely not applicable to the consultation, and therefore could not be classified as inaccurate. Treating them as maximally accurate prevented artificial deflation of overall accuracy scores. Each patient transcript was scored out of 18. The frequency of each classification under each section was calculated (e.g., the presenting complaint was accurate in 3 out of 6 patients). Additionally, the total frequency of each classification was calculated (e.g., 5 sections were classified as Inaccurate WMD). While the approach to evaluating AI-generated documentation was similar to that described by Farooq et al. [2], the human comparator differed, with this study using simultaneous manual note-taking by medical students rather than clinician dictation and medical transcription.

III. Results

Overall, 36 sections were assessed, of which 4 were labelled as N/A. These sections were excluded from analysis as they do not influence the evaluation of Heidi's efficacy, leaving 32 total sections classified (Table 1). The amended classification yielded an overall accuracy rate of 59% (Figure 1). When Inaccurate WNMD sections, which do not affect the overall quality of patient notes, are included, the effective accuracy by section rises to 84%. Despite this, 16% of sections remain inaccurate, highlighting areas where Heidi's notes may require verification (Figure 1).

Each patient note was scored based on the classification of its individual sections, with a maximum possible score of 18 (see Methods). Because patients presented with varying complaints and consultation complexities, this scoring allowed Heidi's efficacy to be assessed in the context of each patient. Patients 5 and 6 both achieved perfect scores, reflecting simpler consultations compared to patients 1–4 (Table 2). Overall, Heidi scored 90 out of a possible 108 points, corresponding to an overall accuracy by patient of 83%, a very similar result (Table 2).

TABLE I: CLASSIFICATION OF 32 SECTIONS

Classification	Number
Accurate	19
Inaccurate WNMD	8
Inaccurate WMD	5
N/A	4
Total	36
Total excluding N/A	32

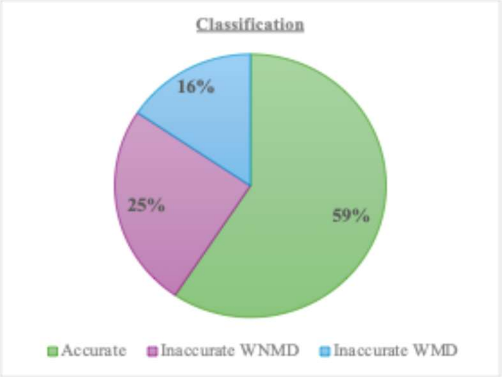

Figure 1: Pie chart showing percentage by classification

ISBN: 978-1-68558-362-0

TABLE II: POINTS SCORED PER PATIENT

Patient Note	Score
Patient 1	13
Patient 2	14
Patient 3	13
Patient 4	14
Patient 5	18
Patient 6	18
Overall Score	90
Total Possible Points	108

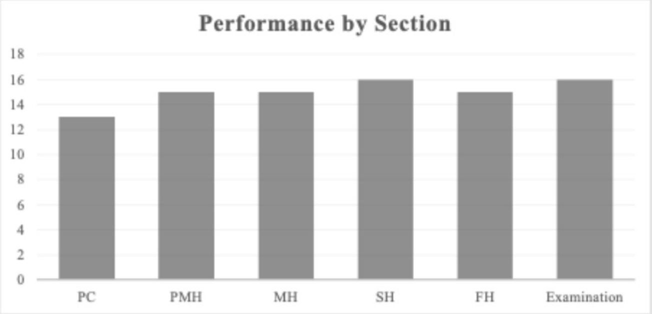

Figure 2: Bar chart demonstrating numerical performance by section

TABLE III: CLASSIFICATION OF PRESENTING COMPLAINT

PC	Number of Sections
Accurate	3
Inaccurate WNMD	1
Inaccurate WMD	2
N/A	0

TABLE IV: CLASSIFICATION OF PAST MEDICAL HISTORY

PMH	Number of Sections
Accurate	3
Inaccurate WNMD	3
Inaccurate WMD	0
N/A	0

TABLE V: CLASSIFICATION OF MEDICATION HISTORY

MH	Number of Sections
Accurate	4
Inaccurate WNMD	1
Inaccurate WMD	1
N/A	0

TABLE VI: CLASSIFICATION OF SOCIAL HISTORY

SH	Number of Sections
Accurate	4
Inaccurate WNMD	2
Inaccurate WMD	0
N/A	0

TABLE VII: CLASSIFICATION OF FAMILY HISTORY

FH	Number of Sections
Accurate	3
Inaccurate WNMD	1
Inaccurate WMD	1
N/A	1

TABLE VIII: CLASSIFICATION OF EXAMINATION

Examination	Number of Sections
Accurate	2
Inaccurate WNMD	0
Inaccurate WMD	1
N/A	3

Each section (as outlined in Methods) was scored according to the number of classifications received across the six patient notes (Accurate, Inaccurate WNMD, etc.). Using the same scoring system, each section was assigned a maximum score of 18, with each section repeating six times for six patients and being scored a maximum of 3 points (Table 2). (Tables 3-8) portray accuracy by section (e.g., presenting complaint). Comparison of these scores highlighted the Presenting Complaint (PC) as the poorest performing section, with a score of 13 (Figure 2). The remaining sections performed similarly, each scoring 15 or 16 points, indicating consistency in Heidi's accuracies and inaccuracies.

IV. DISCUSSION

Looking at the data that was collected, Heidi performed consistently better than initial expectations. Large Language Models (LLM) had been garnering attention for their hallucinatory tendencies and how they frequently produced incorrect or fictional information while generating text [4]. Heidi, as an LLM, was expected to exhibit some of the hallucinatory tendencies seen in other non-healthcare specific LLMs. This was one of the main concerns explored, as inaccurate information in patient notes poses a direct risk to patient safety. According to NHS England, "missing, inaccurate, or non-standard information can lead to inconsistent care and risk the quality and safety of care delivered" [3]. If inaccuracies in Heidi's generated notes reach a significant threshold, this raises concerns that attempts to improve productivity and reduce administrative burden may inadvertently compromise patient safety.

During transcription, we observed one instance of Heidi 'hallucinating' where it recorded that a patient had a sister instead of a brother. This was classified as 'Inaccurate WNMD', as although the detail was incorrect; it did not directly affect patient care. This raises the question of whether such inaccuracies may affect patient confidence. Patients now routinely access their consultation records via the 'Patient Access' and 'NHS' applications. Although inaccuracies within the WNMD section do not directly influence clinical management, they may nevertheless shape patients' perceptions of the overall quality of care received. Such perceptions have the potential to alter the dynamics of

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

the doctor–patient relationship, where even subtle erosions of trust could contribute to suboptimal patient engagement and, consequently, less favourable long-term outcomes.

Overall, Heidi's outputs were consistently accurate or very close to accurate, with only 5 out of 36 sections formally classified as inaccurate and making a difference. It raises the question as to what is considered an acceptable standard in health. When combined with a professional or student performing a final review before integration into patient records, these findings suggest that Heidi could substantially reduce the administrative burden while maintaining accuracy and ensuring patient errors are not made.

When analysing scores by patient transcript, it was observed that Patients 1–4 presented with more complex conditions, either as first-time presentations or due to extensive past medical histories. In contrast, Patients 5 and 6 represented comparatively 'simpler' consultations, without complex comorbidities or the need for extensive physical examinations. Heidi performed at least on par with, and in some cases better than, the student's notes for Patients 5 and 6. The lowest scores were seen in Patients 1 and 3, each receiving 13/18, which aligned with the greater complexity of their cases.

During the review of the independent analyses, occasional disagreements arose regarding whether certain points should be classified as 'Inaccurate WMD' or as 'Inaccurate WNMD'. Because the overall evaluation focused on Heidi's accuracy in the context of patient safety, the final decision consistently leaned toward the more conservative option by classifying points as 'Inaccurate WMD' when there was uncertainty. The category of 'Inaccurate WNMD' was used for minor errors or irrelevant details that a student would not typically record. Considerable discussion centered on what degree of inaccuracy constitutes a genuine patient safety concern. Consequently, when calculating final accuracy percentages, sections labelled as 'Inaccurate WNMD' were not regarded as posing a direct risk to patient safety and were therefore included in the overall accuracy score.

From a medical education perspective, Heidi can help students focus more on developing the soft skills of patient engagement, while also providing a secondary reference against which they can check their own notes during placement. Nevertheless, this raises an important concern regarding how to ensure that clinicians consistently conduct thorough reviews and whether insufficient oversight could inadvertently increase risks to patient safety. Addressing this question lies beyond the scope of the present study but provides scope for further research.

While Heidi exceeded expectations in this study, it comes with its limitations. As with any AI input system, Heidi could not interpret when a patient would point to body parts to explain their symptoms. Heidi transcribed statements verbatim without contextual interpretation on certain occasions. For instance, if a patient said they had a 'dry cough' but produced 'green mucous', Heidi would state the same in the notes. However, a dry cough by definition does not produce mucus. Additionally, Heidi had limited capability to correct information. It frequently stated the first thing a patient said, even if they changed their description throughout the duration of the consultation. Heidi's notes described negative signs and symptoms that a patient had been asked about (no fever, no chest pain) while notes taken by healthcare professionals generally only include positive signs and symptoms unless negative symptoms are diagnostically relevant. While these may be accurate, the question arises of whether recording too much could lead to an information overload where important details are overlooked.

Our study has several limitations. We had a small sample size of six patients, which limits statistical power and the ability to generate findings or test for significance. A reviewer bias may prevail over the descriptive analysis carried out, as whether a point will or will not make a difference can be subject to interpretation. The reviewers were medical students, so judgments about clinical significance may differ from experienced practitioners. Natural variation in typing speed and computer proficiency is expected between medical students, as would be expected between different general practitioners. Additionally, medical students were aware that their documentation would be evaluated, which may have introduced a Hawthorne effect, potentially inflating note quality relative to routine practice. Other factors, such as usability and clinician/patient satisfaction were not measured. This study was only carried out in a general practice setting, but further research should be carried out in different medical settings where outcomes may differ. This study is biased towards one AI tool. Finally, our findings relate to the version of Heidi used during the study. As it is a continually updated software, updates may change accuracy and performance

V. Conclusion and Future Work

All things considered, Heidi achieved an impressive accuracy rate of 84%. While complete accuracy may be an ambitious expectation, even small inaccuracies carry potential implications for patient safety. This reinforces the necessity (emphasised by Heidi itself upon login) that all generated notes must be reviewed before integration into patient records. Heidi functions effectively as a supplement to human documentation but is not yet capable of replacing a general practitioner. Its main strength lies in reducing administrative burden: by generating draft notes, it allows clinicians - or medical students, in this case, to concentrate more fully on patient interaction during consultations, with their role shifting to reviewing and correcting the output. Further research needs to be carried out with larger sample sizes and in a variety of settings to gain more clarity on this subject. It could place attention on long-term effect studies to evaluate how medical scribe use influences clinician burnout, efficiency, and patient outcomes over months or years.

Larger multi-centre studies would generate stronger evidence of long-term advantages and possible drawbacks of AI medical scribes. This trial was not blinded, but future iterations could address this by using practices that routinely record calls and document consultations manually. These recordings and notes could then be randomly selected for review, helping to minimise perfectionism bias that may arise when clinicians or students are aware that their documentation is being assessed. As well as this, comparative studies should be carried out to compare Heidi against other AI-driven scribes, such as ChatGPT, DeepScribeAI, Nuance DAX, Suki AI, Abridge, evaluating which system is the most effective in terms of efficiency, accuracy, and user satisfaction. Given that AI medical scribes vary in their intended purpose, from primary care workflow support to cost efficiency, and clinician explainability, further studies should have evaluation frameworks that measure outcomes aligned with the scribe's intended use. These studies could compare their use in a general practice setting versus a hospital setting, as well as comparison across different specialities, including psychiatry, paediatrics and surgery, to assess whether effectiveness differs by clinical context. It's important that future research investigates cost-benefit analyses to assess if the increase in efficiency from an AI medical scribe outweighs the implementation costs. The wider financial impact of adopting the system at a regional or national scale should also be determined. This would provide crucial data to optimise resource allocation and support decisions on large-scale implementation.

Research should evaluate the impact of AI medical scribes on both patient and staff experience. Regarding patients' experience, patient satisfaction, communication, and privacy concerns should be explored. Notably, patients attending for mental health–related consultations were less willing to participate, highlighting important considerations regarding patient attitudes toward the use of AI in clinical settings. In reference to staff experience, clinician burnout is often linked to excessive administrative workload, so further studies can demonstrate whether the use of AI scribes can reduce administrative burden and therefore improve clinician wellbeing. The current landscape underscores the need for robust evidence on the safety, efficiency, and real-world performance of large language models, as highlighted by the NHS [3], before their broader deployment in medical documentation. Given the rapid advancement of digital technologies, this study aims to evaluate whether such tools can enhance clinical efficiency by reallocating clinician time toward direct patient care. This question is increasingly pertinent as AI becomes more accessible and further embedded within healthcare systems. Another important area of research is the effectiveness of the integration of Heidi with different Electronic Health Records (EHR), such as Epic and EMIS, as smooth integration is needed for its practical use. Such studies should investigate whether integrating Heidi with EHR systems would impact the accuracy of documenting records and operational workflow.

Occasional errors, such as missing medications from the patient's history, could lead to incomplete or inaccurate records and pose risks to patient safety. This does not exclude the possibility that Heidi may eventually achieve full accuracy; however, in its current form, the system still necessitates a degree of clinician oversight, as acknowledged by the software itself. From a psychological standpoint, increasing software accuracy may paradoxically reduce the robustness of final human review over time, as clinicians become less vigilant when errors appear infrequent. It could therefore be argued that the software is, in some respects, safer when errors are more common, as this heightened error frequency prompts greater clinician alertness. Nevertheless, an inaccurate scribe ultimately presents a significant safety concern. Implementing structured safety checks similar to those embedded in electronic prescribing systems. may provide a viable mitigation strategy.

In conclusion, this study adds to the growing body of research demonstrating the promise of AI-assisted clinical documentation. Ensuring that AI scribes genuinely enhance patient care, support clinician workflow, and maintain high standards of clinical governance will be critical to their responsible integration into routine practice.

Acknowledgments

Ethical approval was not required for this small-scale service evaluation. All patients provided informed written or verbal consent for participation and the use of their consultation data. This research did not receive funding from any agency in the public, commercial or not-for-profit sectors. The authors thank the general practice staff in North Central London for facilitating data collection, and the patients who generously consented to participate. The authors would also like to express their gratitude to their supervisor, Dr. James Rusius, for his guidance and oversight during the study.

Author contributions

Conceptualisation: SS; Methodology: SS, AS, CH; Validation: SS, AS, CH; Investigation (data collection): SS, CH; Data curation: SS, AS, CH; Formal analysis: SS, AS, CH; Writing - Original Draft: SS, AS, CH; Writing – Formatting and Editing: SS, AS. All authors reviewed and approved the final manuscript.

References

[1] Heidi Health, "About Heidi Health – Company Story | UK." *Heidi Health*, 2019. [Online]. [retrieved: August, 2025] Available: https://www.heidihealth.com/uk/about-us/company

[2] F. Farooq, H. Cooper, A. Shipman, and C. D. Michell, "Artificial intelligence verses traditional method in generating dermatology consultation letters: a pilot study comparing accuracy, readability, and efficiency," *Clin. Exp. Dermatol.* 2025. [Online]. [retrieved: August, 2025] doi: https://doi.org/10.1093/ced/llaf323,

[3] NHS England, "High quality patient records," *NHS England*, 2022. [Online]. [retrieved: August, 2025]

ISBN: 978-1-68558-362-0

Available: https://www.england.nhs.uk/long-read/high-quality-patient-records/

[4] Y. Sun, D. Sheng, Z. Zhou, and Y. Wu, “AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content,” *Humanit. Soc. Sci. Commun.*, vol. 11, no. 1, 2024. [Online]. [retrieved: August, 2025] doi: https://doi.org/10.1057/s41599-024-038

Copyright (c) IARIA, 2026. ISBN: 978-1-68558-362-0