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

Shreya Shah, Aarya Shetye, Carol Habib
Guy's, King's and St Thomas' School of Medical Education
King's College London
London, United Kingdom
Contact email: shreya.2.shah@kcl.ac.uk

Presenters' Background

Shreya Shah

A third-year medical student at King's College London interested in surgery and obstetrics and gynaecology.

She is interested and involved in academic and clinical research. She also actively serves within society committees, acting as a student representative for King's Health Partners and the Vice President of the KCL Women in Surgery society.

Aarya Shetye

A third-year medical student at King's College London with a strong interest in critical care and respiratory medicine.

Alongside her studies, she is involved in both clinical research and medical education. She previously facilitated play sessions at Evelina London Children's Hospital and plays an active role in organising academic conferences at university, with two years' experience serving on KCL Paediatrics and KCL Anaesthetics societies.

Carol Habib

A third-year medical student at King's College London with a particular interest in anaesthetics and surgery.

She is the Events Lead of the KCL Emergency Medicine Society and has undertaken other lead roles within multiple university society committees, leading the planning and delivery of various events and conferences. She is also actively involved in clinical research.

Introduction

Clinical documentation is essential

- Patient safety
- Continuity of care
- Legal compliance

The documentation burden in primary care

- Time-intensive
- Interrupts doctor–patient communication
- Increasing administrative workload
- Risk of documentation errors

Background and Aim

- Prominence of AI in healthcare: medical scribes
- What is Heidi?
  - Medical scribe that uses artificial intelligence and natural language processing
- Real time speech to structured notes
- What is the current evidence?
  - Limited real-world validation

Aim of this study

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.

Methods

- Free version of Heidi
- ‘General Practitioner’s note’ template
- Patient consultations spanned over three months
- Six consultations
  - five in-person
  - one telephone
- Consultations carried out by two medical students at a time
- Three minutes at the end to collate notes on computer
- Heidi was activated up until general practitioner review
- Compared AI generated notes vs medical student generated notes

Methods

- Three medical students independently reviewed the consultation notes
- Each transcript was scored under a 3-2-1 system, with a total possible score of 18

Each consultation was divided into six sections:

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

Classifications used:

- Accurate (3 points)
- Inaccurate, Will Not Make a Difference (Inaccurate WNMD) (2 points)
- Inaccurate, Will Make a Difference (Inaccurate, WMD) (1 point)
- Not Applicable (N/A) (3 points)

Results

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

Table 1: Frequency of each classification across all consultation sections.

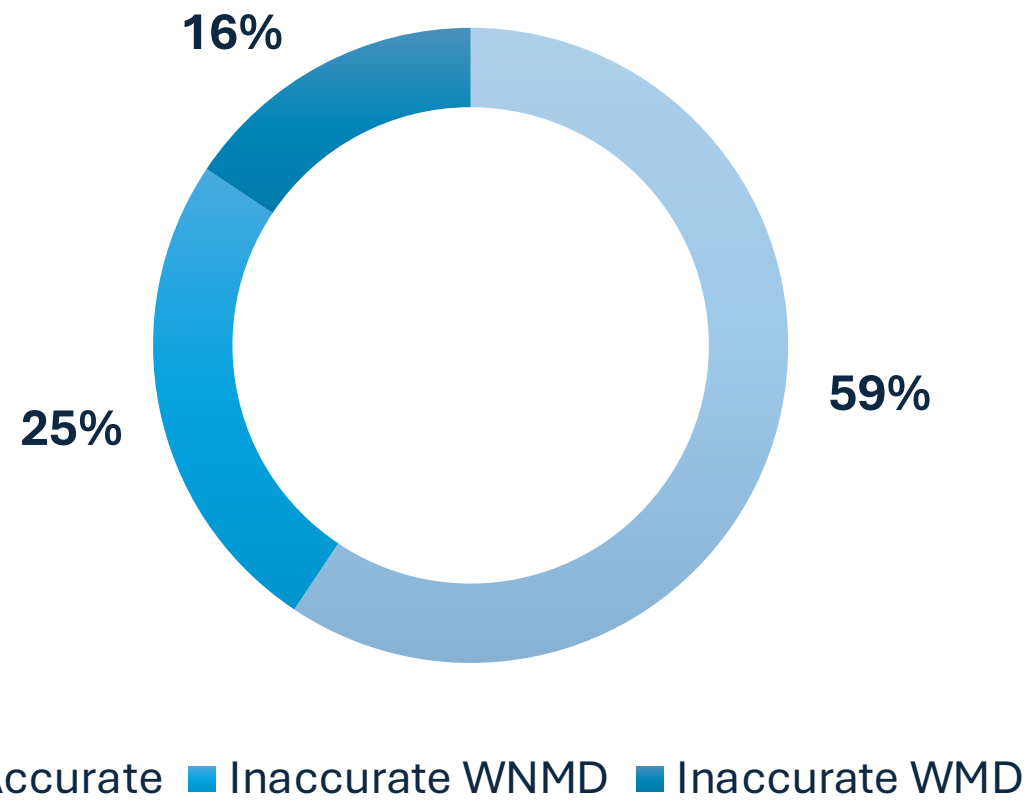

Figure 1: Percentage distribution of accuracy.

Results

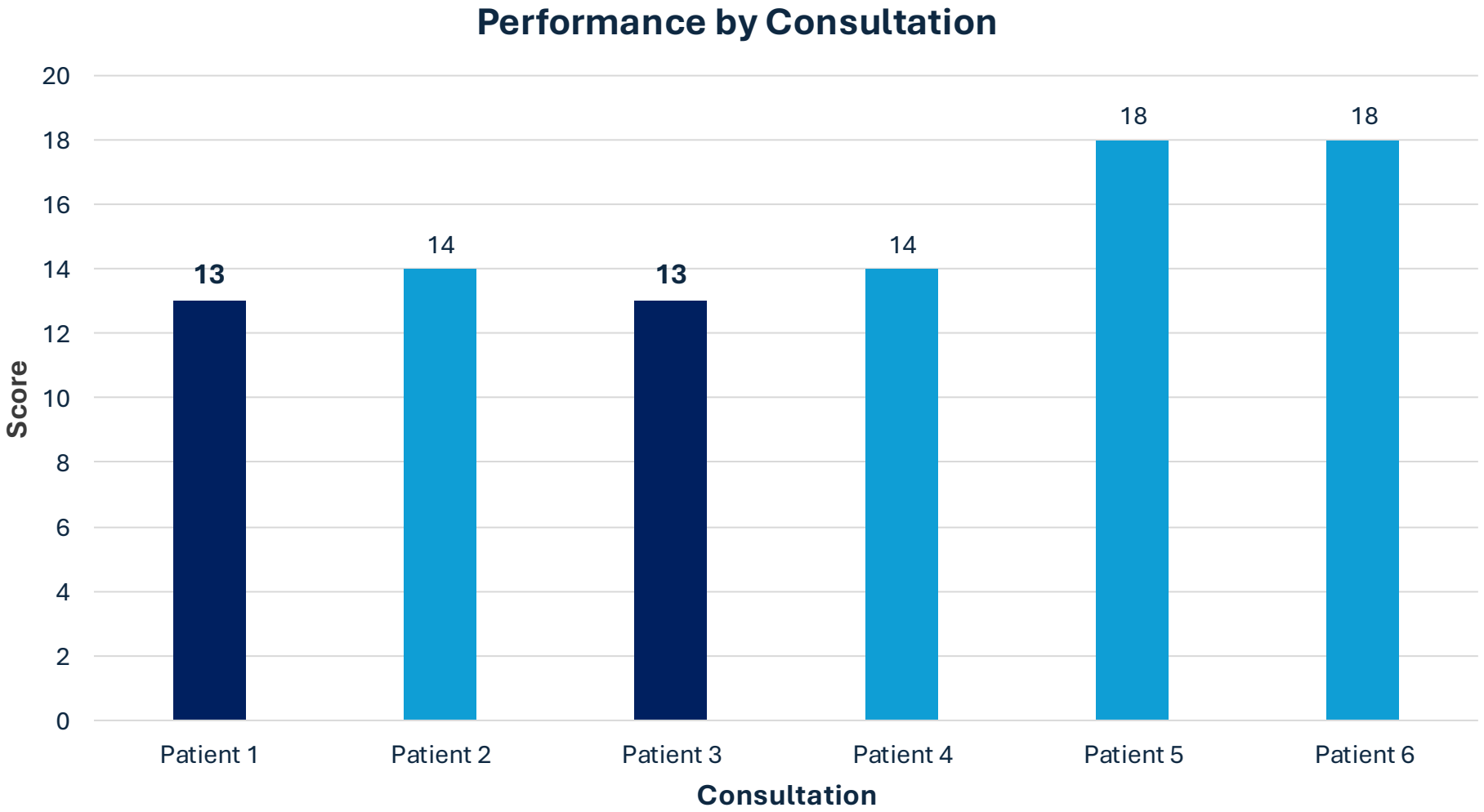

Figure 2: Heidi's performance broken down by individual patient consultation.

Results

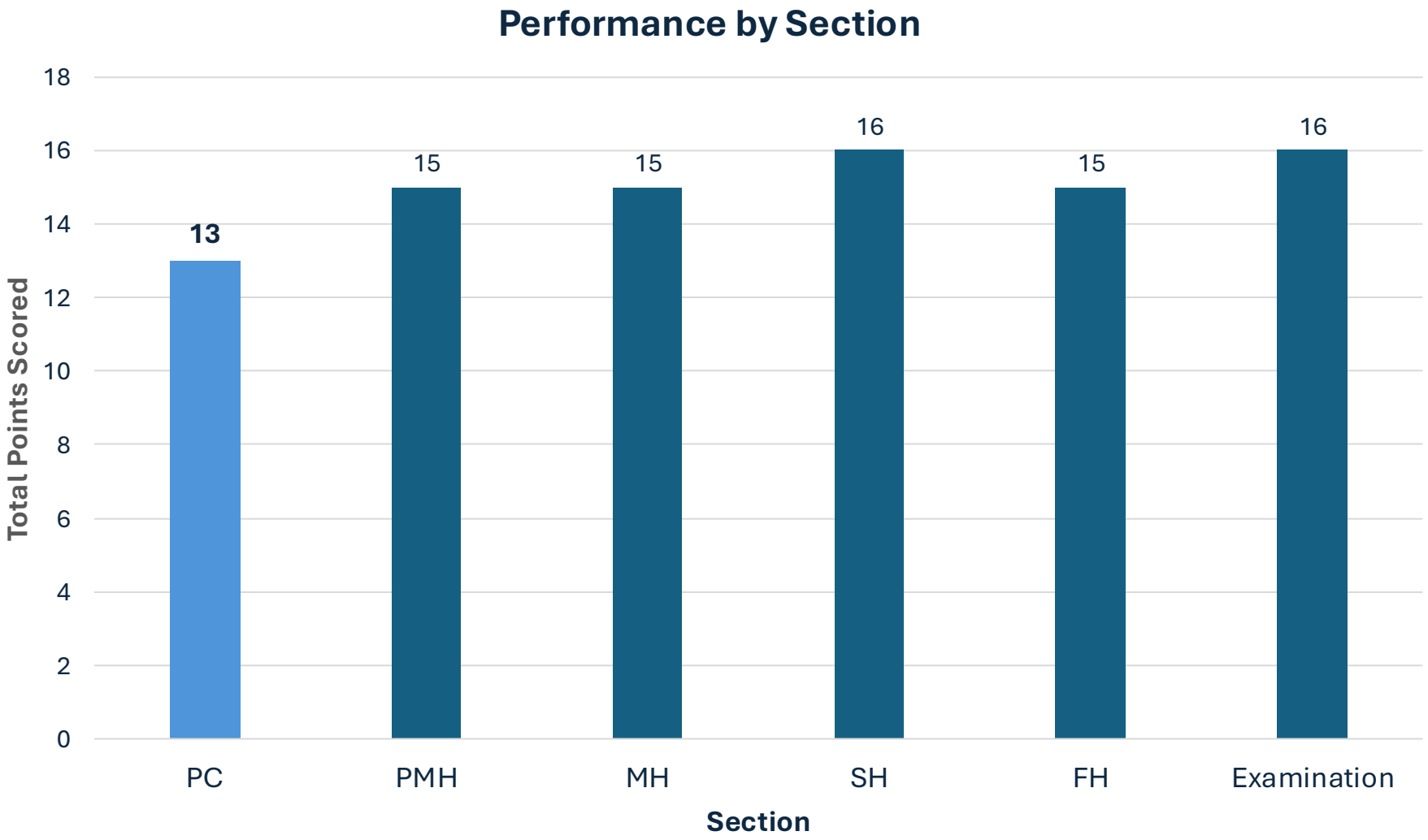

Figure 3: Heidi's performance broken down by consultation section.

So, what do our results suggest?

Performance is promising

- Exceeded initial expectations
- Single hallucination event observed

Emerging system-level considerations

- Defining thresholds for acceptable inaccuracy
- Implications to doctor-patient relationship with systems such as the NHS app and Patient Access

Oversight remains essential

- Not yet independent of reviewer input
- Accuracy declines with increasing case complexity

Limitations

Our study

- Small sample size
- Reviewer bias
  - Subjective
  - Medical student reviewers
- Variation in speed and computer proficiency
- Hawthorne effect
- Not measured:
  - Clinician/patient satisfaction
  - Varied settings
  - Other AI tools
- Heidi continually updated

Heidi

- Missed non-verbal symptom cues
- Verbatim transcription without contextual interpretation
- Limited capability to correct information
- Described both positive and negative signs and symptoms

What can we learn from this?

- Clinically significant errors were rare
- Errors mainly minor omissions/context issues
- However, even small inaccuracies matter in healthcare
- Human review remains essential
- Functions as a supplement, not a replacement
- Important to consider risk of human review becoming less robust over time

Heidi's Accuracy: 84%

Future directions

Academic

- Larger, multi-centre validation studies
- Long-term outcomes (burnout, efficiency, safety)
- Comparative studies across AI scribes
- Comparative studies across different specialties

Implementation

- Cost-effectiveness and system-level impact
- Evaluation of patient and staff experiences and attitudes
- Integration with electronic health record systems

Key takeaways

- High accuracy is pivotal in healthcare documentation
- Human oversight is still needed
- Responsible integration is possible

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: doi: https://doi.org/10.1093/ced/llaf323, August, 2025]
- [3] NHS England, “High quality patient records,” NHS England, 2022. [Online]. [retrieved: August, 2025] 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, https://doi.org/10.1057/s41599-024-038

Thank you! Any questions?

shreya.2.shah@kcl.ac.uk

aarya.shetye@kcl.ac.uk

carol.habib@kcl.ac.uk