Evaluation of Artificial Intelligence (AI) Scribes in Medical Practice: Cross-Regional Analysis

Dr Anju Soni[1] and Dr Ian Treasaden[2]

[1]Canberra Health Services, Canberra, Australia and [2]West London NHS Trust, London, United Kingdom

doi: 10.1192/bjo.2025.10184

Aims: This study aimed to evaluate the implementation of AI scribes in medical practice across Australia and England, focusing on assessing current adoption rates, measuring awareness of perceived benefits and potential risks and developing recommendations for safe and effective deployment.

Methods: A comprehensive survey was conducted across 50 medical practitioners equally distributed between Australia and England. Practitioners were from both urban (60%) and rural (40%) environments, with a representation across experience levels: junior (30%), mid-career (45%), and senior (25%) practitioners. Data was collected over 4 weeks through a semi-structured questionnaire.

Results: The analysis revealed significant insights into AI scribe adoption and perception across both countries. Of the surveyed practitioners, 28% (14) are currently utilising AI scribes in their practice, while nearly half (48%) are actively considering implementation. The remaining 24% expressed no immediate plans to adopt this technology. Reported AI scribe benefits were notably high among respondents, with time-saving potential being the most recognised advantage (90% awareness). Practitioners demonstrated strong recognition of the technology's ability to reduce administrative burden (84%) and improve patient interaction (76%). However, the assessment of documentation quality improvements was lower at 62%.

Risk awareness varied significantly across different aspects. Privacy concerns dominated the risk perception landscape, with 78% of practitioners expressing awareness of potential privacy issues. Clinical accuracy risks and legal liability concerns were acknowledged by 70% and 64% of responders respectively. A crucial finding was that only 42% of practitioners were aware of their medical defence union's position on AI scribe usage, revealing a significant knowledge gap in professional liability coverage.

Among current users, satisfaction levels showed a mixed picture. While 64% reported positive experiences (21% very satisfied, 43% somewhat satisfied), a notable portion remained neutral (22%) or expressed dissatisfaction (14%). Implementation concerns centred primarily around training requirements (80%) and system integration challenges (72%), with medical defence coverage emerging as a significant concern (62%).

Conclusion: The study highlights the critical need for healthcare providers to establish comprehensive implementation strategies that address both technical and legal considerations. Practitioners in both regions must prioritise verification of medical defence union coverage before adopting AI scribes. UK medical defence unions have clearer guidelines compared with Australian medical defence organisations. Australian practitioners should align their implementation with RACGP digital health guidelines, while UK practitioners need to ensure NHS Digital compliance.

The findings emphasize that successful AI scribe implementation requires a balanced approach that addresses technical integration, risk management, and insurance coverage. The high level of interest, coupled with significant uncertainty about medical defence coverage, indicates a clear need for professional organisations to provide more detailed guidance on this emerging technology.

Abstracts were reviewed by the RCPsych Academic Faculty rather than by the standard *BJPsych Open* peer review process and should not be quoted as peer-reviewed by *BJPsych Open* in any subsequent publication.

LUMEN: Prototype Conversational AI to Streamline Dementia Assessments

Dr Song Ling Tang[1], Mr Alexander Robertson[2], Dr Huizhi Liang[2], Prof John-Paul Taylor[2] and Dr Judith Harrison[2]

[1]Hertfordshire Partnership University NHS Foundation Trust, Hertfordshire, United Kingdom and [2]Newcastle University, Newcastle, United Kingdom

doi: 10.1192/bjo.2025.10185

Aims: Dementia assessments are time-intensive and often distressing for patients and caregivers. Underdiagnosis of non-Alzheimer's disease subtypes remains prevalent. This study aimed to develop and evaluate LUMEN (Large Language Model for Understanding and Monitoring Elderly Neurocognition), a prototype conversational AI to automate caregiver-collateral data collection before clinical appointments. Our goals were to reduce clinician time per assessment, improve diagnostic accuracy across dementia subtypes, and standardise caregiver assessments.

Methods: LUMEN's development integrated a Patient, Public, and Professional Involvement (PPPI) process, incorporating stakeholder workshops, a modified Delphi process with 130 clinicians, and iterative consultations to identify key diagnostic priorities, such as functional impairments, safety concerns, and inclusivity. Four open-source 7B-parameter large language models (LLMs) – Mistral, Llama2, Zephyr, and Phi2 – were evaluated for efficiency (token count), readability (Flesch Reading Ease), and contextual relevance (cosine similarity to clinical dialogues). Mistral:7B was selected and fine-tuned using automated hyperparameter adjustments (GridSearchCV), advanced prompt engineering (chain-of-thought, flipped classroom techniques), and BLEU-scored linguistic refinement. A prototype interface was tested using 16 clinician-simulated caregiver dialogues derived from case vignettes spanning dementia subtypes and normal cognition. LUMEN's diagnostic outputs were compared with clinician-derived diagnoses using the Area Under the Receiver Operating Characteristic (AUROC) curve and agreement measured via Cohen's kappa. Usability was assessed via the System Usability Scale (SUS).

Results: LUMEN demonstrated strong performance in distinguishing dementia from normal cognition (AUROC=0.89) but moderate subtype differentiation (AUROC=0.66). Agreement between LUMEN and clinician evaluations was substantial (Cohen's κ=0.82). However, Lewy body dementia (DLB) identification lagged due to symptom-reporting inaccuracies. System Usability Scale (SUS) scores (mean=82/100) exceeded the 'excellent' threshold ($\geq$80). PPPI feedback highlighted LUMEN's potential to standardise assessment and reduce waiting times.

Conclusion: LUMEN is a promising conversational AI tool for improving dementia diagnostics. Gathering caregivers' collateral input before appointments could streamline workflows within existing outpatient systems and improve clinical accuracy. Real-world trials would help assess workflow integration and mitigate vignette-based biases from simulated testing, such as the overrepresentation of typical phenotypes.

This study was conducted in collaboration with Mr Bede Burston, Dr Elizabeth Robertson, and Dr Donncha Mullin, whose contributions were invaluable.

Abstracts were reviewed by the RCPsych Academic Faculty rather than by the standard *BJPsych Open* peer review process and should not be quoted as peer-reviewed by *BJPsych Open* in any subsequent publication.