BMJ Health & Care Informatics

Cognitive science in the evaluation of medical AI systems

Vimla Lodhia Patel [1,2]

To cite: Patel VL. Cognitive science in the evaluation of medical AI systems. *BMJ Health Care Inform* 2023;**30**:e100929. doi:10.1136/bmjhci-2023-100929

Received 06 October 2023
Accepted 04 December 2023

[1]Cognitive Studies in Medicine and Public Health, New York Academy of Medicine, New York, New York, USA
[2]Biomedical Informatics, Columbia University Vagelos College of Physicians and Surgeons, New York, New York, USA

Correspondence to
Dr Vimla Lodhia Patel; vpatel@nyam.org

Clinical cognition is central to a clinician's daily tasks, such as making diagnostic and therapeutic decisions. For example, doctors rely on their memory to recall relevant facts, concepts and experiences that can help them diagnose and treat their patients. Memory is needed for clinicians to accumulate and update their evidence-based knowledge from prior cases.[1] Similarly, doctors perceive and make decisions through observations of the physical and mental state of their patient. Their ability to sense the patient's mood, emotions or personality clearly plays an important role.

One of the critical components of a doctor's cognitive task is higher-level clinical reasoning required to analyse and synthesise the information that they gather from various sources (such as history, physical examination, laboratory tests and imaging data). They use deductive reasoning to apply principles to specific cases. Similarly, they use inductive reasoning to infer general principles from specific cases they have seen. Abductive reasoning, where deduction and induction are intermixed, is often used in a natural clinical setting, supporting the generation of hypotheses or explanations based on incomplete data.[2] Of course, a physician's thinking process is also prone to errors and biases that can affect the quality and safety of healthcare. Therefore, physicians need to be aware of their cognitive strengths and limitations and must seek ways to improve their skills to overcome cognitive challenges. Decision-support systems, such as those using artificial intelligence (AI) methods, can augment and support clinicians to alleviate some of the problems. How should these AI agents interact with clinicians in the clinical world, and what evaluations are required to assure that these systems are efficient, effective and safe? (Descriptions of such detailed evaluation methods have been published elsewhere.[3])

The field of human-computer interaction intersects cognitive-behavioural, computer, and information sciences. As healthcare systems become more sophisticated and intelligent, careful evaluation of these tools, as they are actually leveraged by intended users, becomes necessary.[4] Human-machine dyads too often end up on the technology-led rather than the human-led side.[5] The implementations often fail to support physicians in their tasks, highlighting system inadequacies and demonstrating why human-centred approaches to designing and evaluating AI tools are even more critical. The human-centred AI strategic framework is appropriate for evaluation because it understands technology as a tool to empower, augment and enhance human agency instead of emulating or competing with it.[5]

Applied medical AI and medical cognition mutually influence each other in several ways, including providing a basis for developing formal models of clinical competence in problem-solving tasks. An essential publication that significantly influenced the field of clinical cognition is the 1972 classic *Human Problem Solving* by Newell and Simon,[6] where human problem-solving was explicitly linked to research in AI. The theoretical framework provided in this volume offered a language for the study of cognition. It introduced protocol analysis, a set of dominant methods used in investigations of high-level cognition such as comprehension and reasoning.

In order to evaluate the impact of intelligent systems on human reasoning and thinking, a technique known as *verbal think-aloud* (or simply *think aloud*) is often used to capture rich descriptive data on the thought processes that underlie human actions.[7] The authors who popularised this approach specified the conditions under which verbal reports are acceptable as legitimate data. My colleagues and I have undertaken several studies using verbal think-aloud methods to investigate the nature of reasoning using clinical systems, including the associated effects of expertise and decision-making skills.[2 8]

During the think-aloud process, the subjects' statements, revealing what they are thinking as they do their clinical tasks, are audio-recorded, transcribed and analysed using methods of natural language coding. Due to misunderstanding this process, some think-aloud data have been collected retrospectively, where the subject can reconstruct the information in memory (with potential for memory distortion). This retrospective approach leads to insights and explanations that are considered suspect. More appropriately, think-aloud protocols that collect observational data in context, while the subject is actually solving the problem, provide richer data for characterising cognitive processes. The generated verbal data are usually referred to as a 'protocol' and may then be subjected to *protocol analysis*.[7]

Qualitative evaluation techniques, where clinicians functioning as users are involved in the assessment process, are often leveraged in naturalistic field studies within the context of dynamic clinical workflow. When new or unknown technologies create challenges for users, people instinctively turn to see if there are technological solutions to the problem they have encountered. However, these challenges that arise in AI systems often cannot be mitigated through technical means alone. If the user turns to solutions that do not include broader clinical and societal insight, their approach may only compound the system's dangers since the technology, no matter how well it functions in a laboratory, continues to struggle to function optimally in the real world. Evaluation design must accordingly be targeted in the context of the broader sociotechnical systems in which such assessment is always embedded.[9] Besides technological and cognitive factors, these include understanding the sociocultural and organisational structures of the environment and of the community at large. A sociotechnical approach avoids any structural imbalances, providing opportunities for a broader participation to consider diversity, such as race and ethnicity. Sociotechnical AI safety redistributes power from a single group to a broader, diverse community.

If one sees the future of AI as a way of working together with intelligent human beings, then the concept of *augmented* intelligence is also a vital consideration. Appropriate evaluation will create opportunities to improve the design of clinical AI systems to ensure clinician control while leveraging the latest technological developments to increase automation. In a 2022 Berkley AI research blog,[10] Miao and Liu introduce the concept of *a human-machine loop* where humans and machines are mutually augmenting each other. One can argue that such loops exist in real-world clinical applications. Instead of replacing clinicians' intelligence, augmented intelligence envisions using AI methods in an assistive role.[11]

This change in emphasis has broad implications for evaluation. Technologies mediate clinicians' performance and influence how they behave as they interact with them; these systems enhance clinicians' ability to perform tasks better and change how they do such tasks. Cognitively based evaluation to understand higher-level thinking and reasoning is necessary to capture the precise nature of such change and to offer optimal means to intervene. Human beings and technologies, including AI systems, are different in nature, even though machines can mimic some aspects of human behaviour. Human beings have unique qualities and weaknesses that set them apart from machines. Our challenge is to leverage both optimally while understanding their strengths and encouraging relevant synergies while guarding against over-reliance on either extreme.

Twitter Vimla Lodhia Patel @VimlaLPatel

Competing interests None declared.

Ethics approval Not applicable.

Provenance and peer review Commissioned; externally peer reviewed.

ORCID iD

Vimla Lodhia Patel http://orcid.org/0000-0003-1656-6642

REFERENCES

1 Wallin A, Kettunen P, Johansson PM, *et al*. Cognitive medicine: a new approach in health care science. *BMC Psychiatry* 2018;18:42.
2 Patel VL, Kaufman DR, Kannampallil TG. Diagnostic reasoning and expertise in health care. In: *The Oxford Handbook of Expertise*. New York, NY: Oxford University Press, 2019.
3 Shortliffe EH, Sepùlveda MJ, Patel VL. Framework for the evaluation of clinical AI systems [Cognitive Informatics in Biomedicine and Healthcare]. In: Cohen TA, Patel VL, Shortliffe EH, eds. *Intelligent Systems in Medicine and Health: The Role of AI [Internet]*. Cham: Springer International Publishing, 2022: 479–503.
4 Patel VL, Kaufman DR, Kannampallil T. Human-computer interaction, usability, and Workflow. In: Shortliffe EH, Cimino JJ, eds. *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*. Cham: Springer International Publishing, 2021: 153–75.
5 Shneiderman B. In: *Human-Centered AI*. New York: Oxford University Press,
6 Newell A, Simon HA. *Human Problem Solving*Reprint ed. edition. Brattleboro, Vermont: Echo Point Books & Media, 2019: 938.
7 Ericsson KA, Simon HA. Protocol analysis. In: *Protocol Analysis: Verbal Reports as Data*. Revised Edition. Cambridge, MA, USA: A Bradford Book, 1993.
8 Patel VL, Groen G. The general and specific nature of medical expertise: A critical look. In: Ericsson KA, Smith J, eds. *Toward a General Theory of Expertise: Prospects and Limits*. 1st edition. Carlsbad, CA: Cambridge University Press, 1991: 93–125.
9 Patel VL, Shortliffe EH, Stefanelli M, *et al*. The coming of age of artificial intelligence in medicine. *Artif Intell Med* 2009;46:5–17.
10 Miao Z, Liu Z. Rethinking human-in-the-loop for artificial augmented intelligence [Internet]. The Berkeley artificial intelligence research Blog. Available: http://bair.berkeley.edu/blog/2022/05/03/human-in-the-loop/ [Accessed 5 Oct 2023].
11 Patel VL, Cohen TA. Clinical cognition and AI: from emulation to Symbiosis [Cognitive Informatics in Biomedicine and Healthcare]. In: Cohen TA, Patel VL, Shortliffe EH, eds. *Intelligent Systems in Medicine and Health: The Role of AI [Internet]*. Cham: Springer International Publishing, 2022: 109–33.