mage source: Images 2.0 / The Rundown

Beyond Doctor versus AI: The Case for Structured Clinical Intelligence

"Old AI model tops doctors in ER trial" is an arresting headline. It is also slightly misleading, not because the result is unimportant, but because the deeper lesson is not a simple contest between machines and clinicians. The real question is no longer whether AI can produce impressive clinical reasoning from written information. It clearly can. The more important question is how that capability should be joined safely to clinical expertise.

A recent *Science* paper by Brodeur and colleagues tested OpenAI's o1/o1-preview model across a series of demanding clinical reasoning tasks. These included NEJM clinicopathological cases, NEJM Healer cases, Grey Matters management cases, landmark diagnostic cases, probabilistic reasoning tasks, and 76 real emergency-room cases drawn from electronic health-record text. In the emergency-room experiment, the AI generated more accurate differential diagnoses than two attending physicians at three clinical touchpoints: initial triage, ER physician assessment, and admission to the hospital or intensive care unit. It did this despite working only from written record data, without seeing, hearing, examining, or speaking with the patient.

The result is striking for another reason. o1-preview was not the latest possible model. In AI terms, it was already an older generation system. Yet it performed at or above expert human level on several narrow, text-based diagnostic and management reasoning tasks. In the ER cases, o1 achieved exact or very close diagnostic scores of 67.1% at initial triage, 72.4% at the ER physician encounter, and 81.6% at admission, exceeding both physician comparators at each stage.

It would be tempting to read this as evidence that AI is now "better than doctors." That would be the wrong conclusion. What the paper shows is more precise and more interesting: current frontier AI is exceptionally strong at a particular kind of work — written clinical reasoning over structured or semi-structured clinical information.

That distinction matters. A large language model is operating in its native medium when it reads a written case and generates a written answer. It can hold a long case history in context, retrieve a wide range of diagnostic possibilities, generate a broad differential diagnosis, avoid fatigue, and produce a polished response that aligns well with formal scoring rubrics. It is very good at asking, in effect, “What else could this be?”

Human clinicians are doing something different. They are not merely solving written diagnostic puzzles. They are examining patients, noticing distress, interpreting tone, movement, appearance and interaction, communicating with families, negotiating uncertainty, prioritising action, and working within the constraints of time, beds, staffing, tests, patient preference, and organisational pressure. They also carry responsibility for the consequences of action. These are not minor additions to clinical reasoning; they are part of clinical reasoning.

The authors of the paper recognise this boundary. They note that their study addressed only text-based performance and that real clinical medicine includes non-textual inputs such as the patient’s level of distress, visual information, and imaging interpretation. They also warn that benchmarks based on carefully curated or cleaned-up cases may overstate AI performance when compared with the messier information available in real workflows.

This is why the “team” question is crucial. The decisive issue is not doctor-alone versus AI-alone. It is whether a properly designed doctor-plus-AI system can outperform both the unaided clinician and the unaided model in ways that matter clinically.

The paper does not fully answer that question. It contains some relevant comparator data involving physicians with GPT-4, and those results are sobering. In the Grey Matters management cases, o1-preview scored a median of 89%, whereas GPT-4 scored 42%, physicians with GPT-4 scored 41%, and physicians using conventional resources scored 34%. In the landmark diagnostic cases, o1-preview scored 97%, GPT-4 scored 92%, physicians with GPT-4 scored 76%, and physicians with conventional resources scored 74%.

Those findings should make us pause. Simply putting AI in front of a clinician does not automatically create a better clinical system. The phrase “human in the loop” can be deceptively reassuring. A poorly designed loop may add little. The clinician may ignore the AI, defer to it too readily, be anchored by its first suggestion, or consult it too late to change the course of reasoning. Human-plus-AI is not automatically additive. It has to be designed.

A proper team study would need to compare at least four conditions: clinician alone, AI alone, clinician using AI freely, and clinician using AI in a structured workflow. The structured workflow is the most important. The clinician should first make an independent judgement. The AI should then generate a differential diagnosis, identify cannot-miss alternatives, suggest missing information, challenge the working diagnosis, and flag management risks. The clinician should then make the final decision, documenting whether and how the AI changed the plan.

This would test something more clinically meaningful than benchmark accuracy. It would examine whether AI can reduce premature closure, widen the diagnostic frame, improve recognition of rare but dangerous conditions, calibrate confidence, strengthen management planning, and reduce unsafe omissions. In prospective clinical trials, the outcomes should go further still: delay to correct diagnosis, adverse events, escalation, revisits, unnecessary testing, length of stay, cost, workload, and clinician trust.

My expectation is therefore deliberately nuanced. Doctor-plus-AI should beat doctor-alone, especially where the danger is missed diagnosis, cognitive narrowing, incomplete differential diagnosis, or failure to consider rare but serious possibilities. AI is likely to become a powerful second-opinion engine and diagnostic safety net.

At the same time, AI-alone may continue to beat doctor-plus-AI on narrow text-based benchmarks. If the task is simply to read a written case and produce a scored differential diagnosis or management plan, the model may remain the strongest performer. That is not surprising. The benchmark rewards complete, explicit, text-based reasoning, and the AI is optimised for exactly that output.

But clinical practice is not a benchmark. The safest and most valuable configuration in real care is likely to be structured doctor-plus-AI. The AI brings breadth, recall, consistency, and tireless textual reasoning. The doctor brings embodied assessment, contextual judgement, ethical responsibility, patient communication, practical feasibility, and the ability to integrate non-textual signals. Neither is complete alone.

The clinical prize is not replacement. It is disciplined coupling.

The Brodeur paper should therefore be read not as the end of the doctor, but as the beginning of a new diagnostic work system. The central design question is no longer whether AI can reason. It can. The question now is how human and artificial reasoning should be joined safely, so that the combined system is not merely cleverer, but better for patients.

Reference

Brodeur, P.G., Buckley, T.A., Kanjee, Z., Goh, E., Ling, E.B., Jain, P., Cabral, S., Abdulnour, R.-E., Haimovich, A.D., Freed, J.A., Olson, A., Morgan, D.J., Hom, J., Gallo, R., McCoy, L.G., Mombini, H., Lucas, C., Fotoohi, M., Gwiazdon, M., Restifo, D., Restrepo, D., Horvitz, E., Chen, J., Manrai, A.K. and Rodman, A. (2026) 'Performance of a large language model on the reasoning tasks of a physician', *Science*, 392(6797). doi: 10.1126/science.adz4433.