

Microsoft's MAI-DxO diagnosis results are published in Science and go to a hospital study

Source copy kept by wheresthe.ai for the story "Microsoft's MAI-DxO diagnosis results are published in Science and go to a hospital study" (<https://wheresthe.ai/d/us-microsoft-mai-dxo-science-paper-hospital-study-2026-10/>).

Source: <https://microsoft.ai/news/building-medical-superintelligence/>

Text saved 23 Oct 2026 (the page as published).

SHA-256 of the saved text: a5c308a64db3fe54ccc3701d9c721b017c4577dab7b3c89385a9e3386e880b22

Copilot -->

Health

Building Medical Superintelligence

MAI Health Team

October 22, 2026

Health

MAI Health Team

LI

X

FB

What if AI could help clinicians work through a complex medical diagnosis, step by step, and achieve higher accuracy, at lower cost, with fewer medical tests?

That future is now coming into view. On common presentations to a US emergency department, MAI-DxO – our AI diagnostic system – is 1.7 times more accurate than physicians when evaluated using cases from MIMIC-IV-ED.

And on medicine's most complex cases, published in The New England Journal of Medicine , MAI-DxO is more than twice as accurate as physicians using textbooks and the internet – and nearly four times as accurate when they diagnose without those resources. It also manages to do this at half or even one third of the cost, getting to the right diagnosis with fewer tests and procedures.

Our work is published today in Science .

Read more about our vision for medical superintelligence also published this week in BMJ Digital Health and AI .

Rethinking How AI is Built and Tested

AI has the potential to transform health and healthcare, and progress is accelerating rapidly. The challenge now is to build and rigorously evaluate AI systems that translate advances in their clinical reasoning capabilities into meaningful impact.

Today's paper, building on earlier work , introduces two foundations for that shift: the Sequential Diagnosis Benchmark (SDBench), which tests step-by-step diagnostic reasoning, and the Microsoft AI Diagnostic Orchestrator (MAI-DxO), our leading AI system on the benchmark.

Medical licensing exams such as the United States Medical Licensing Exams (USMLEs) are designed to assess physicians' knowledge – not the realities of making a clinical diagnosis. In these tests, all information is presented upfront and a single answer is selected; in practice, clinicians must build the picture step by step, deciding what to ask, which tests to order and how each new finding changes the likely diagnosis. Such exams have become the common benchmarks for evaluating the performance of health AI .

That's why we developed SDBench, which evaluates performance across two complementary sets of cases: 304 complex cases taken from The New England Journal of Medicine (NEJM) case proceedings and a set of more typical emergency-department presentations from MIMIC-IV-ED. Each becomes an interactive case in which AI

models and/or physicians start with limited information and choose what to request next. The benchmark measures diagnostic accuracy, evidence requested and estimated test costs – assessing both the answer and the path taken.

We built MAI-DxO and tested it within SDBench. Rather than relying on one model for an immediate answer, it operates like a panel of virtual physicians: each proposing diagnoses, selecting useful questions or tests, and challenging the emerging view. MAI-DxO coordinates this exchange over several rounds of debate, updating its assessment as evidence arrives and weighing each test's likely value against its cost.

What we found:

On complex NEJM cases, MAI-DxO achieved 83% diagnostic accuracy. Physicians achieved 22% without external reference resources and 36% with them, making MAI-DxO about 3.7 and 2.3 times more accurate, respectively.

On more typical emergency-department cases from MIMIC-IV-ED, MAI-DxO achieved 80% accuracy versus 47% for physicians – more than 1.7 times higher.

When optimised more heavily for cost, MAI-DxO performance remained strong while choosing more cost-effective workups. On NEJM cases, one configuration achieved 81% accuracy at less than half the cost of comparative systems.

Across both datasets, no other human or AI approach was both more accurate and less costly than MAI-DxO.

These gains persisted – and improved – when we repeated the evaluation with newer models, including GPT-5.5, without changing the orchestration method, suggesting the approach is not tied to one model generation.

We are open-sourcing SDBench for select NEJM cases – including the benchmark, Gatekeeper, Judge and cost-estimation code – so researchers can reproduce the evaluation, test other models, and challenge or extend the findings.

Explore SDBench

From Better Diagnosis to Better Care

Diagnosis is only one part of the care journey. Better care also requires keeping pace with medical knowledge, selecting the right treatment, understanding change over time, and supporting choices that reflect a person's values and access to care. The opportunity is to unite these capabilities into a standard of care not widely available today. In *The Building Blocks of Medical Superintelligence*, published in *BMJ Digital Health and AI*, we explore what it would take for AI to support this broader vision – and ultimately exceed the capabilities of today's strongest clinical teams and health systems.

Sequential diagnosis offers an early example: AI physicians collaborating, challenging assumptions and refining their clinical reasoning as evidence emerges. Our ambition is to build AI that connects knowledge, context and decisions across the care journey – not simply excels at one clinical task. Through our collaboration with Mayo Clinic, we aim to combine frontier AI, clinical expertise and longitudinal multimodal data to build systems for the complexity and diversity of real-world care.

Testing in the Real World

Benchmarks matter, but the true test is performance in real clinical environments. Through a new partnership with Beth Israel Deaconess Medical Center in Boston, we will evaluate clinical reasoning capabilities across different hospital settings. In a prospective study, patients will interact with the system directly, with performance measured against physicians' independent assessments. Working alongside clinical teams will show where the technology succeeds, where it falls short and how it can fit into busy workflows.

Health at Microsoft

Rapid advances in medical AI are opening new possibilities for patients and clinicians. To turn this progress into practical impact, we are bringing together health teams from across Microsoft – including scientists from Microsoft Research and the teams behind products like Dragon Copilot and PowerScribe – into an integrated effort at Microsoft AI. Working together, we will accelerate the delivery of trusted AI systems that drive better clinical outcomes and care experiences.

Build the Future With Us

We're a lean, fast-moving lab made up of some of the world's most talented minds. We have an exciting roadmap of compute at MAI, which is ramping quickly and extensively. And we have an ambitious mission we truly believe in. We're also fortunate to partner with incredible product teams giving our models the chance to reach billions of users and create immense positive impact. If you're a brilliant, highly-ambitious and low ego individual, you'll fit right in—come and join us as we work on our next generation of models!

[Explore all jobs](#)

More Stories

[Health Check: How People Use Copilot for Health](#)

[health](#)

[Our values in operation: Health](#)

[partnerships](#)

[Introducing Copilot Health](#)

[health](#)