AI in Healthcare: "Impressive" Progress, Missing Medico-Economic Proof

Jérôme Vetillard

CTO | VP R&D | Chief Product Officer | AI-Powered Healthcare & Life Sciences Products | Compliance by Design | PhD AgroParisTech | CPO MIT Sloan | Exec MBA IE Business School & Brown University

2 juillet 2025

Last week Microsoft AI unveiled **MAI-DxO**, a multi-agent “orchestrator” that cracked **85 %** of 304 complex *NEJM* cases—four times the score of a control group of physicians and at lower notional cost. The system acts like a virtual panel of specialists that questions, orders tests, and double-checks itself before naming a diagnosis. (The Path to Medical Superintelligence | Microsoft AI)

Why this matters

1. **Beyond multiple-choice** – Sequential, cost-aware reasoning mimics real workflow better than USMLE-style benchmarks that LLMs already dominate.
2. **Cost signal baked in** – Every test carries a CPT price tag, attacking the $100 B/year over-testing problem in the U.S. alone. Being a young resident, you always tend to over-test not to miss the rare diagnosis of your patient (having an obligation of means), so "senior guidance" is always welcome and AI does not sleep during night shift when your senior told you not to awake him/her for nothing :)
3. **Virtual breadth + depth** – One orchestrated agent can span generalist coverage *and* sub-specialist nuance—something no single clinician can do.

Read the fine print

- **Sampling bias** – Rare, teaching-oriented *NEJM* cases don't match everyday primary-care prevalence. These are "signal rich" cases, and if you understand AI as being fundamentaly "signal processing", you know it's always better to have a high signal over noise ratio.
- **Hindsight bias** – Solved cases retro-fitted into dialogues may leak textual cues.
- **Baseline bias** – 21 GPs, offline and outside their specialty, are a fragile yard-stick for "super-human" claims.
- **Model-on-model & sponsor bias** – Microsoft's own LLM grades its sibling; incentives and shared blind spots lurk. Remember "Majorana quantic chipset" now the buzz is over ?
- **Synthetic results** – The Gatekeeper invents lab values when absent; only 0.2 % were human-audited, and regulators have yet to define QA rules. The use of synthetic data is still under investigation by FDA/EMA (especially for control arms for "In Silico" clinical trials).

What's still outside the sandbox (but at the heart of clinical practice)

- **Text-only semiology** – No images, sounds, smells, or tactile signs.
- **"Post-anamnesis" starting point** – A pre-digested vignette bypasses the messiness of real history-taking.
- **No unruly patients** – The AI never confronts inconsistency, emotion, or non-verbal cues (you know, what you get from people named "patients") so bedside transferability remains untested.

Bottom line

The benchmark shows how elegantly an LLM navigates a curated corpus—not how it practices bedside medicine. To **earn clinical credibility**, we still need:

- multimodal inputs (images, **auscultation**, **free-form speech**),
- prospectively collected patient cohorts, and
- **independent evaluation.**

The (always unanswered) medico-economic question

1. **Clinical edge** – Will AI genuinely outperform—or best augment—human clinicians? Does these technologies can be efficiently integrated in the value chain of healthcare providers ?
2. **Economic impact** – Will it cut system-wide costs and improve outcomes, or become **a Trojan horse letting hyperscalers siphon public-health budgets**?

Until these questions are settled, let's celebrate the technical milestone—**and keep demanding REAL WORLD COST AWARE EVIDENCE.**

Medico-economic evaluation must be integrated in every AI for health initiative from its inception.

#AI #Healthcare #DigitalHealth #MedTech #CostEffectiveness #GenerativeAI