Same word, different worlds: AI Terminology in healthcare series

Debate #3: Performance Metrics

Health AI Working Group 2 — AI Evaluation & Validation in Healthcare

Discussion held: June 9, 2026

What is this series about?

When it comes to AI evaluation and validation in healthcare, terms like "Human-in-the-loop," "performance metrics," "bias," or "validation" appear in all kinds of content across sectors. Yet if you sit a clinician, a regulator, a developer, a manufacturer, a procurement officer, and a program evaluator in the same room and ask, "What do you mean by that word?" you will quickly find far more nuance than you might have expected. That's why the Working Group on AI Evaluation and Validation in Healthcare at HealthAI — The Global Agency for Responsible AI in Health, identified the need for a different kind of terminology discussion: built through structured debate with professionals from across countries and sectors, and discussed with HealthAI members leading diverse initiatives across 15 countries. A term may have one agreed meaning, several definitions, or a single meaning that carries important nuances. The idea is to encourage and embrace a shared and diverse understanding of each term across different stakeholder groups which will bode well for transdisciplinarity in this field. Each piece is authored by the speakers of that debate session, and the views expressed reflect their individual perspectives.

Contributors

Roxana Salehi, PhD — Founder & Principal Consultant, Vitus Consulting (Barcelona)

Roxana is a strategic consultant advising healthcare leaders on design, measurement, and evaluation of complex healthcare initiatives around the world. More recently her work has focused on AI-enabled healthcare, including precision mental health. The perspective that she brings to this debate is that of a program designer and program evaluator.

Daniel Futeman — Project Director, Vital Strategies (South Africa)

Daniel has more than 15 years of experience as a technologist building and scaling digital health and AI solutions across Africa. With deep expertise in health systems architecture and digital social and behaviour change systems, and a

track record of bridging AI research and responsible deployment, he brings a practical, evidence-based approach to advancing AI in low- and middle-income country health systems. Daniel currently leads the CODA (Cause of Death Determination Assistant) initiative at Vital Strategies, working with a global consortium to develop and test an AI-powered cause-of-death ascertainment tool designed to strengthen mortality surveillance in LMIC health systems.

Paul Campbell — Chief Regulatory Officer, HealthAI

Dr Paul Campbell brings with him a wealth of expertise in healthcare, information technology, and regulatory frameworks. Prior to his role as Chief Regulatory Officer at HealthAI, Paul served as the Head of Software and AI at the Medicines and Healthcare Products Regulatory Agency (MHRA), where he led various innovative projects aimed at shaping the future of medical device regulation in the UK. Notably, Paul spearheaded key initiatives within the Software and AI Medical Device Change Programme, including the regulation of digital mental health tools and the AI Airlock regulatory sandbox—a groundbreaking approach to addressing the challenges of regulating AI-driven medical devices.

A respected clinician and thought leader, Paul is recognized for his extensive knowledge and experience in healthcare technology regulation. He is also a proud graduate of the NHS Digital Academy's inaugural cohort, further solidifying his role as a pioneer in the digital health landscape.

Rado Andriantsimiavona — Director of AI, HealthAI

Rado Andriantsimiavona spearheads global initiatives at HealthAI to advance AI governance, regulation, and ethical deployment in healthcare.

With over 20 years of experience driving transformative digital health and AI innovations in the Healthcare and Life Sciences sectors, Rado is renowned for his hands-on expertise in digital strategy, solution-building, and forging impactful partnerships. Beyond HealthAI, Rado is a recognized thought leader in the Swiss and international Digital Health Innovation ecosystem, championing Value-Based Healthcare (VBHC) and the "5Ps of Medicine"—Personalized, Predictive, Preventive, Participatory, and Populational.

Moderated by: Luciana Pires - Impact & Strategy Manager, HealthAI

Luciana is an impact management professional with experience in strategy, program design, and fostering collaboration among diverse stakeholders, having worked across Europe and Latin America. At HealthAI, she leads the development of the Community of Practice, building connections between patient advocacy groups, regulators, academics, and innovators. She also heads the organization's impact assessment efforts, drawing on her background in impact consulting.

Key Considerations

1. The definition of performance metric was broadly shared across speakers.

Simply put, a performance metric is a measurable result — quantitative or qualitative — used to assess an AI model, tool, or system in a healthcare context.

2. There was agreement that performance metrics should not focus solely on model performance — but speakers had different frameworks for conceptualising a broader view.

- **Roxana** proposed three factors that determine what constitutes a valid performance metric: (1) the *evaluand* — the thing being evaluated (the AI model, its implementation, or its impact); (2) the *unit of analysis* — whether measurement is at the level of the model, the system, the organisation, or the healthcare system; and (3) the *evaluation lens* — the methodological and epistemological framework determining what counts as evidence, and whose perspective is captured.
- **Rado** noted that the appropriate metrics depend on the type of problem being solved: different problems require different metric families. For example, model evaluation metrics could include classification metrics (e.g. accuracy, F1 score, area under ROC curve), or regression metrics which are used for predicting a number (e.g. Mean Squared Error).

  The "right" choice is however meaningless on its own. Standard statistical metrics only measure performance in an isolated mathematical vacuum. To evaluate how a model will actually perform in the real world, these metrics must be layered with calibration, contextualisation, and localisation; i.e. rooted in real world data. It's only at deployment and through continuous adaptation that the true performance, value, and impact, of an AI system can be evaluated. **Calibration** ensures an AI model's predicted probabilities match real-world outcomes. **Contextualisation** grounds the model's output in relevant facts, user intent, and real-world constraints. **Localisation** adapts an AI model to the specific language, cultural nuances, legal compliance, regional standards, and clinical and socio-technical realities of a particular target market. Last but not least: what about the human-AI system performance? What tools do we set to measure it?
- **Paul** drew on the medical device regulatory framework, discussing clinical association, scientific validity, and technical/analytical performance. The **clinical association** was particularly highlighted as relevant, as it considers

the performance of the product in relation to its intended use. It is vital for AI as SaMD to review and follow the relevant IMDRF and MDCG guidance for Clinical Evaluation. Although somewhat out of date, MEDDEV 2.7/1 rev. 4 remains a primary document to follow for clinical evaluation of software used in healthcare. Ultimately, **evidence is required across the three paradigms of Clinical Association, Analytical Validation and Clinical Evaluation**. Other relevant resources can be found here.

- **Daniel** referenced the FDA definition — "quantitative or qualitative measures that can be used to assess the ability of a model to produce the desired output for a given task. The choice of the metrics depends on the specific task and the model objectives." He also talked about using the *AI lifecycle* to organize how we talk about performance metrics: model development, first deployment, adoption, monitoring, and post-market evaluation all require different measures. That point was also mentioned by Rado.

3. A broad perspective on performance metrics does not mean measuring everything. It means choosing the right metrics for the specific problem we are trying to solve in a specific clinical and organizational context.

There was strong agreement that the goal of performance measurement is not to measure everything but rather to measure the right things. Attempting to measure everything is not practical and misses the point. The goal should be selecting metrics that are meaningful for the specific context and stage in which the AI is operating. They should also be feasible for teams to collect, and tied to a clear 'change pathway' (how do we think our actions in this AI project lead to real positive change in the world?)

4. Performance metrics are not purely technical choices; they reflect priorities, values, resource constraints, risk tolerance, and trade-offs.

All speakers had this reflection, referencing different examples:

- Sometimes improvements in accuracy may increase computational cost, or introduce new risks.
- Metrics that are about equity and inclusion, for example, can show the underlying values associated with an AI project or product. An example given was an accountability metric, such as the percentage of project budget allocated to patient co-design sessions.
- Different stakeholders may prioritise different performance dimensions, such as clinical utility, cost, workflow, or adoption. However, **safety** and **risk** provide an essential shared starting point: what risks does the tool introduce, what measures allow those risks to be identified, what mitigations are in place, and what residual risks remain. In this sense,

risk-related metrics may provide a common foundation before expanding to broader performance dimensions.

- Acceptable performance thresholds also depend on the risk profile of the use case. A model achieving 70% accuracy may be appropriate where a human reviews every output, but inadequate where errors carry clinical or operational consequence and oversight is limited. Risk characterisation shapes which metrics matter and what counts as good enough.

5. The real meaning of each performance metric is context-specific and depends on what's considered 'good enough'

The TRICORDER study was referenced as a concrete illustration: the AI assessed performed 2.3 times better than clinicians at detecting relevant clinical signs — yet showed no measurable improvement in clinical outcomes. This means that model performance metrics did not translate into clinical value. So can we say the model performance metrics were inherently 'good'? There was agreement in the group that they should be analyzed in relation to other relevant metrics such as usability for clinicians.

Performance metrics also require clarity about what they are being interpreted against. Depending on the stage of evaluation and the type of AI system, the relevant reference point may be a reference standard, benchmark dataset, baseline model, human expert judgement, clinical guideline, current workflow, accepted safety threshold, or prior real-world performance. Without this context, a metric such as "90% accuracy" is difficult to interpret. The question highlighted was "*90% accuracy against what*? *A perfect dataset?*". Here, the discussion also highlighted the importance of real-world data, not only training data, to be used in such metrics.

6. Program evaluation — which extends beyond randomised controlled trials — offers useful frameworks for determining meaningful metrics in complex healthcare settings.

Theory-based evaluation (e.g., theory of change approach, results-based management) is one such lens. There is growing recognition from the Medical Research Council and the National Institute for Health Research of the value of theory-based approaches to evaluation in complex settings. This approach examines not only whether an AI model/tool works, but why it works, how it interacts with the context in which it is implemented, and how the evidence it generates can support decision-making, learning, and iterative improvement. Metrics under this lens include contextual factors that influence AI implementation, adoption, and impact, such as baseline organisational conditions before AI introduction (e.g., staff attitudes toward the tool, existing

challenges with electronic health records, the regulatory landscape already in place or approaching).

7. "Cultural translators" — people who can bridge technical, clinical, procurement, and governance discussions of measurement — are critical.

Different stakeholder groups use the same terms to mean different things, or different terms to mean the same thing. Technologists, clinicians, procurement officers, and governance bodies all privilege different metrics and bring different expectations to what performance means. Having someone or a team who can facilitate the translation, surface the "aha moments," and ensure alignment can be very helpful.

Sources

European Commission (2016). MEDDEV 2.7/1 revision 4: Clinical evaluation: a guide for manufacturers and notified bodies under Directives 93/42/EEC and 90/385/EEC. https://webgate.ec.europa.eu/circabc-ewpp/rest/download/7423d348-1e94-4fb8-8a65-02dd2075dec0

European Commission, Directorate-General for Health and Food Safety (n.d.). Guidance: MDCG endorsed documents and other guidance (section: Clinical investigation and evaluation, performance studies and evaluation). https://health.ec.europa.eu/medical-devices-sector/new-regulations/guidance-mdcg-endorsed-documents-and-other-guidance_en

International Medical Device Regulators Forum, Software as a Medical Device Working Group (2017). Software as a Medical Device (SaMD): Clinical Evaluation (IMDRF/SaMD WG/N41FINAL:2017). IMDRF. https://www.imdrf.org/documents/software-medical-device-samd-clinical-evaluation

International Organization for Standardization & International Electrotechnical Commission (2022). ISO/IEC 22989:2022. Information technology: Artificial intelligence: Artificial intelligence concepts and terminology. https://www.iso.org/standard/74296.html

Kelshiker, M., Bächtiger, P., Petri, C., et al. (2026). Triple cardiovascular disease detection with an artificial intelligence-enabled stethoscope (TRICORDER) in the UK: a cluster-randomised controlled implementation trial. The Lancet, 407, 704-715.

https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(25)02156-7/fulltext

Salehi, R. (2026). AI evaluation in healthcare: a decision-analytic tool for selecting meaningful performance metrics. Vitus. https://www.vitushq.com/en/blog/ai-evaluation-in-healthcare-a-decision-analytic-tool-for-selecting-meaningful-performance-metrics/

Skivington, K., Matthews, L., Simpson, S. A., Craig, P., Baird, J., Blazeby, J. M., & Moore, L. (2021). A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ, 374, n2061. https://doi.org/10.1136/bmj.n2061

U.S. Food and Drug Administration (n.d.). FDA Digital Health and Artificial Intelligence Glossary: Educational Resource. https://www.fda.gov/science-research/artificial-intelligence-and-medical-products/fda-digital-health-and-artificial-intelligence-glossary-educational-resource#6a3997a1d06da

Acknowledgement

We would like to thank our members: Robab Abdolkhani (Monash Partners Academic Health Science Centre), Mohamed Baraka (Academic Institute in United Arab Emirates), Estela Bastos (Visual Thinking), Bernardo Blanco (Noxus), Aryan Chaudhary (BioTech Sphere Research), Mona Duggal (ICMR), Khaled El Iskandarani (Global Health Institute - American University of Beirut), Jeremy El Tan (MoH Singapore), Camilo Erazo (Movement Health Foundation), Daniel Futerman (Reach Digital Health), Kristine Cantin Garside (J&J), Jane Hirst (The George Institute for Global Health, Imperial College London), Thomas Kangah (HealthTech4Africa), Diana Khasanova (Brainphone LLC.), Gumindu Kulatunga (Health Information Unit, MoH), Claude Mbekalo Mambo (TechSkills Academy), Luciana Pires (HealthAI, The Global Agency for Responsible AI in Health), Johanan Dravium Ponniah (MoH Singapore), Roxana Salehi (Vitus Consulting), Shubhanan Upadhyay (Consultant), Dijana Vilic (Guy's and St Thomas' NHS Foundation Trust), Nicole Zuk (Siemens Healthineers).