



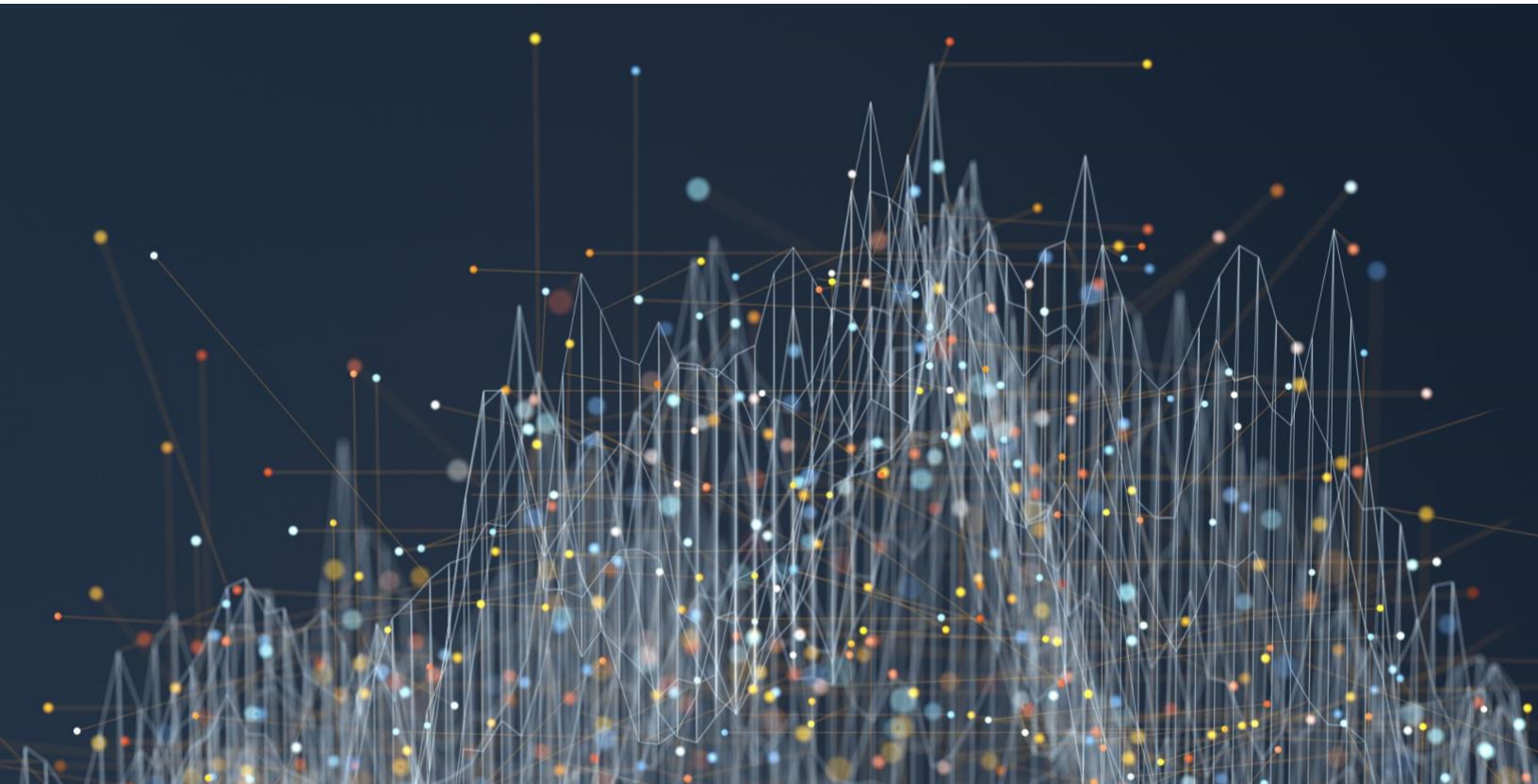
Vega Health White Paper

AI Model Evaluations Build Trust & Drive Outcomes in Healthcare

Trust – and adoption – of AI in healthcare workflows is a complex challenge. Pre-deployment model evaluation provides health systems, clinicians, patients, developers, and regulators with reliable ways to validate AI performance and safety before it impacts care delivery.

October 5, 2026 | Insights Lab

By Chris Provan, Joanne Kim, Aden Klein, and Mark Sendak



01 Why Model Evaluation Matters

Care delivery includes more technology-augmented processes than ever before. Health systems increasingly provide clinicians and patients with access to AI tools spanning scheduling automation to predictive and generative decision support.

The teams implementing these AI products and the regulators deciding on the commercial availability of these AI products need a way to assess the value and performance of an AI product before committing to costly implementation and change management processes. Once implemented, end-users need to understand how the tool works, trust (and verify) its outputs, and responsibly apply appropriate guardrails.

As health systems embrace AI products to improve patient outcomes and lower the cost of care, earning the trust of clinicians and patients is critical to driving use case adoption and sustained value. Despite the widespread use of AI in clinical workflows, trust remains low.¹

Similarly, regulators are increasingly faced with opportunities to grant AI products market access. Given that healthcare is widely seen as the industry vertical with the greatest potential to benefit from AI, many developers and regulators are keen to harness AI to improve public health. These high-stakes decisions need to be grounded in high-quality evidence that independently verifies the performance of AI technologies on real world data.

Pre-deployment evaluation is critical to bridging the gap between AI innovation and the measurable impact of AI technologies. Through model evaluations, health system and regulatory decision makers can be equipped with robust real-world or clinical validation evidence, verify model performance for local patient populations, and increase transparency to build trust with clinicians, patients, and the public.

Developers can leverage evaluations during the product design process to establish parameters for acceptable performance, identify opportunities to improve product performance, build safety guardrails to mitigate risks uncovered during evaluations, and identify what end-user education is needed to promote effective use.

Health systems can use the results of pre-deployment evaluations to generate patient-specific clinical validation evidence, support AI lifecycle management and governance decisions, and build trust with end-users. Evaluations also highlight the context where the technology should not be used, how end-users will act on model outputs, and what measures can help deliver better care outcomes.

End-users – whether clinicians, operators, or patients – can rely on evidence from a rigorous evaluation to understand how an AI technology may affect their workflows or interactions. Evaluations introduce transparency into how an AI model can be a valuable tool and how to most appropriately exercise individual judgement of outputs.

Regulators can use results from independent model evaluations to determine whether a model can improve outcomes in a specific healthcare setting and what safety measures need to be in place to protect patients and the public. Evaluations also help characterize failure modes and inform guidelines for responsible use.

¹ Sofia Guerra and Steve Kraus, “The 2026 AI ROI Scorecard.” Bessemer Venture Partners, October 2026. <https://www.bvp.com/atlas/the-2026-healthcare-ai-roi-scorecard>

Case Study: The Role of Model Evaluations in Utah’s AI Regulatory Sandbox

Approaches to regulating the use of AI in healthcare vary widely from state to state. The uncertainty introduced by a rapidly evolving landscape makes it harder for members of the public – as well as health systems and providers – to confidently purchase and use these transformative technologies.

Policymakers must strike a balance between building robust regulatory guardrails to prevent patient harm and reducing barriers to the adoption of emerging technologies that can improve access, quality, and cost of care. In Utah, the Office of AI Policy (OAIP) has taken a novel approach to addressing this.

Utah’s AI Regulatory Sandbox provides regulatory relief to innovative AI technologies following submission (and approval) of robust evidence of safety and efficacy, a deployment plan with appropriate safeguards, and regular post-deployment monitoring data.²

In order to validate the components of these applications, Utah OAIP has identified trusted third-party evaluators to help conduct and validate pre-deployment evaluation, review deployment and safety guardrails, and conduct post-deployment monitoring.

Vega Health’s pre-deployment evaluation capabilities will help provide evidence to inform these decisions, building trust and advancing positive outcomes from the use of AI in healthcare.

02 Designing an Evaluation

Model evaluation is not one-size-fits-all. To achieve meaningful outcomes from AI implementation, evaluation methodology must flexibly incorporate the insights most relevant to each product’s workflow, model dimensions, and intended end-users.

For an evaluation that primarily supports health system decision makers, the priority is local validation of the AI technology on patient encounters that represent the context of use. Vega Health works with the health system to understand the intended use, implementation context, and the desired outcomes associated with use. The evaluation is then conducted on local historical data to generate evidence supporting prospective use.

For an evaluation that primarily supports regulators, the priority is independent validation of the AI technology on an independent dataset. The developer and the regulator agree on a study design and analysis plan that generates data to support regulatory decisions. Vega Health then works with the AI developer to conduct the study with a partner site using real-world data.

Once the objectives of the evaluation are set by health system or regulatory decision makers, a study design and analysis plan is confirmed. The study design can include analyses across multiple model dimensions. Exhibit 2 includes example measures across domains.

² Alice Schwarze and Zach Boyd, “Evidentiary Expectations for Healthcare AI Sandbox.” Utah Department of Commerce Office of AI Policy, August 2026. <https://commerce.utah.gov/wp-content/uploads/2026/08/Evidentiary-Expectations-for-Healthcare-AI-Sandbox.pdf>

Exhibit 2. Example Model Evaluation Dimensions

	Model Dimension	Evaluation Implications
01 Workflow	Clinical Operational	Determines applicable workflows, end-users, and actions
02 Modality	Predictive Generative	Determines which evaluation metrics best apply
03 Output	Continuous Threshold Free Text	Determines how outputs are presented and evaluated
04 Decision	Inform Advise Direct Automate	Determines how end-users act on model outputs
05 Risk & Harm	False Positive False Negative	Determines where harm can occur from model outputs
06 Feedback Loop	Yes No	Determines if model use changes the ability to evaluate outputs

Notably, the study design and analysis plan is tailored to the AI technology and intended use case. A generative model used to summarize encounters cannot be evaluated in the same way as a predictive model that identifies patients at risk of deterioration.

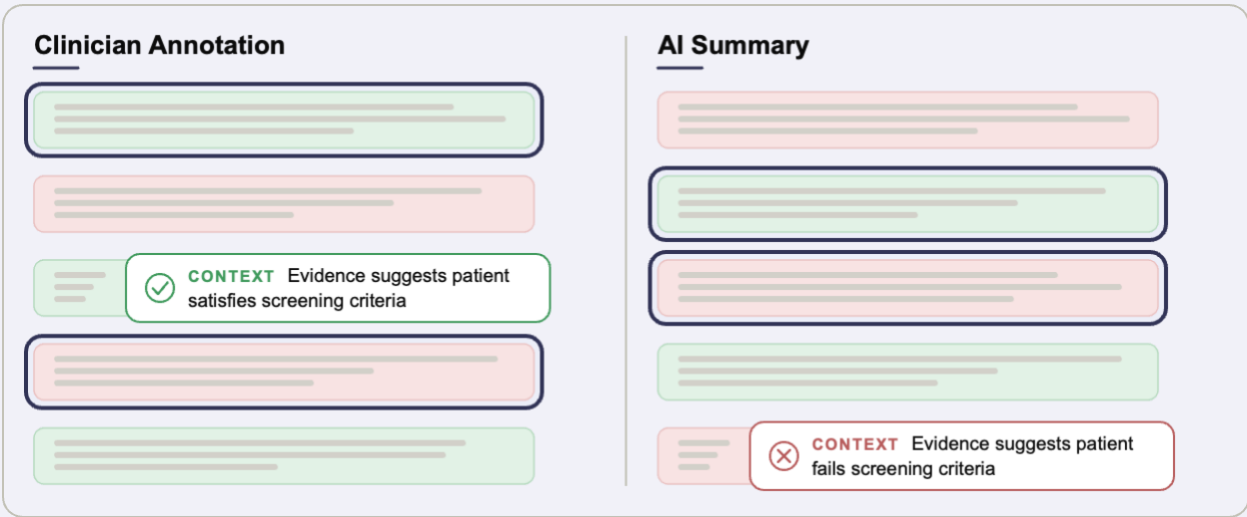
Vega Health maintains a library of evaluation metrics, tests, and statistical analyses that accelerate and streamline AI model evaluations. While the exact combination of analyses may vary for different AI products, each evaluation is intended to generate evidence of model statistical performance, workflow implications (e.g., alert burden, false positive rate), potential risks, bias, and achieved outcomes.

Case Study: Evaluation of a Generative Patient Screening Product

Vega Health conducted a gold-standard evaluation of a generative AI product that produces LLM-generated summaries of patient fit and eligibility for a hospital program based on a set of clinical criteria. The model generates summaries of key evidence from patient charts and overall program eligibility scores to support clinician decision-making.

To evaluate this model, Vega Health compared model outputs to independent clinician annotations and decisions across a sample of patients. The evaluation compared inter-clinician agreement rates, model-clinician agreement rates, and the share of clinician-cited evidence included in model chart summaries. These insights enabled health system leaders to determine how well the model identified relevant chart information and replicated clinicians’ reasoning.

Exhibit 3. Reasoning & Evidence Recall Analysis Mechanism



03 Stages of Evaluation

From initial review for potential implementation to deployment, model evaluations should be repeated across three stages. Each stage has a specific purpose, with specific statistical analyses and target metrics.

Stage 1: Retrospective Evaluation. Vega Health enables real-world evaluation and benchmarking of AI models using large historical patient cohorts. In this stage, AI technologies are tested on a dataset of historical patient encounters from within the target population across a selected period (e.g., 12-24 months). Retrospective evaluations offer an opportunity to evaluate AI products on a large sample of patient encounters across patient subgroups (e.g., race/ethnicity, age, insurance status, level of care), generating robust evidence for technical performance and clinical or operational outcomes based on what model outputs likely would have been had the model been in use at that time.

For generative models, specifically, retrospective evaluations may include the comparison of LLM-generated summaries or recommendations against clinician-generated outputs (“gold standard evaluation”) or by another LLM (“LLM-as-judge”). This methodology can produce a rubric-based score of overall and dimension-specific AI product performance.

This stage of model evaluation is typically conducted with the goal of efficiently generating statistical evidence of model performance over a large cohort of patients with diverse representation. Retrospective evaluations are particularly helpful for generating reliable evidence of output quality, clinical equivalence, potential harms, model bias, and accuracy statistics.

Stage 2: Silent Evaluation. Silent evaluations generate evidence of model performance on real-time clinical encounters or operational cases. While methodology may vary by model type, silent evaluations involve running an AI model without exposing those outputs to users in a way that could influence decisions or outcomes. Model outputs are logged and can be compared with unassisted decisions and outcomes to gain confidence in agreement rates and identify bias or risks. This stage is particularly important for AI products that need to be integrated into end-user workflows, including a focus on how end-users may interact with and act on model outputs.

The real-time clinical environment and data pipelines are materially different than in retrospective analyses in ways that can impact model performance. Key model inputs data may be delayed in the live environment, reducing the accuracy or timeliness of predictions. Clinical notes may still be in draft form, altering the information available to a generative LLM. The timing of model outputs relative to when clinical decisions must be made may point out issues that need to be resolved before live tool use. The key objectives of a silent evaluation are to verify that real-time model performance is comparable to retrospective performance, establish baselines that live model monitoring will be compared against, and inform workflow integration.

Stage 3: Limited-Scope Pilot. Pilots are the final step in evaluating the safety, reliability, and impact of AI products prior to a full rollout. A pilot is the first opportunity to observe and evaluate the actual impact of an AI product on workflows and outcomes. Pilots deploy a tool to a subset of potential users over a defined period of time for live, real-world use. Evaluations focus on comparing workflows and outcomes among pilot users against pre-pilot baselines and against comparable non-participants over the same pilot timeline. Pilots do not have to rise to the standard of randomized controlled trials to provide valuable insights supporting model performance, risk mitigation, and estimates of impact and return on investment (ROI).

Pilots should be scoped to generate the desired insights and support risk mitigation and AI governance requirements. Pilots can require significant investment and involve collaboration and investment of time from end-users. Pilots should therefore only be launched after sufficient retrospective and/or silent trial evaluations have been conducted to minimize the risk of impacting end-user trust and/or time.

04 Building Trust & Driving Outcomes

AI developers, health systems, clinicians, patients, and regulators face increasingly complex decisions about how to build, implement, and monitor AI in care delivery. Without the right clinical validation, local performance data, and operational insights, it becomes incredibly difficult to build workflows and guardrails that really improve outcomes.

Pre-deployment evaluation is a missing lever that enables developers to build products responsibly, health systems to deploy AI confidently, and regulators to prioritize AI products best positioned to benefit the public. Through rigorous AI model evaluations, Vega Health makes it easier to build trust in the performance, safety, and value of AI technologies deployed to improve healthcare.

About the Authors

Chris Provan leads AI Evaluation & Monitoring for Vega Health. **Joanne Kim**, Director of Product & Services, helps Vega Health partners leverage AI investments to drive clinical and operational value. **Aden Klein** leads Vega Health's implementation engagements. **Mark Sendak** is the Co-Founder and CEO of Vega Health.