

CIRCLE DATASETS FOR HEALTHCARE AI REFERENCE STANDARDS

September 2026

TABLE OF CONTENTS

EXECUTIVE SUMMARY	3
HEALTHCARE AI IS ALREADY EVERYWHERE.....	4
HOW THE LAYERS CONNECT.....	5
<i>The First Mile Is Now Machine-Written</i>	<i>5</i>
<i>Decisions Are Increasingly Automated End To End</i>	<i>6</i>
<i>The Question Every Layer Now Faces</i>	<i>6</i>
THE EVALUATION GAP	6
CLEARED DEVICES ARE RARELY TESTED ON OUTCOMES.....	6
EARLY RECALLS ARE ASSOCIATED WITH MISSING VALIDATION	6
MODELS DEGRADE WHEN THEY LEAVE THE DEVELOPMENT SETTING.....	7
THE LABEL DEFINES WHAT A MODEL LEARNS.....	7
GENERATIVE AI ADDS ERRORS AT THE SOURCE	7
SUMMARY	8
WHY CURRENT REFERENCES FAIL	8
NO MODEL CAN BE VERIFIED AGAINST ITS OWN INPUTS	8
PUBLIC BENCHMARKS ARE CONTAMINATED AND SATURATED	8
EXPERT OPINION IS MOST VARIABLE WHERE AI IS USED MOST	9
THE RECORD REFLECTS BILLING, COPYING AND OTHER MODELS.....	9
SYNTHETIC DATA INHERITS ITS SOURCE.....	10
WHAT THE FAILURES HAVE IN COMMON	10
WHAT A REFERENCE STANDARD FOR CLINICAL AI MUST BE.....	11
TEN REQUIREMENTS	11
WHAT THE FRAMEWORKS SAY	11
WHO REQUIRES OUTCOME-ANCHORED EVALUATION	12
REGULATORS	13
HEALTH SYSTEMS	13
PAYERS AND STATES.....	14
ENFORCEMENT AND LITIGATION.....	14
INSURERS.....	15
AI DEVELOPERS.....	15

CIRCLE DATASETS AS EVALUATION DATASETS	16
HOW CIRCLE DATASETS ARE PRODUCED	16
PROPERTIES THAT MEET THE REQUIREMENTS	17
THREE PROPERTIES NO OTHER REFERENCE COMBINES.....	17
DELIVERY OPTIONS	18
WHAT CIRCLE DATASETS CAN AND CANNOT TEST	18
TASKS WHICH CIRCLE DATASETS ARE BUILT TO EVALUATE.....	18
HOW A CASE BECOMES AN EVALUATION ITEM.....	19
TASKS WHICH CIRCLE DATASETS DO NOT ADDRESS.....	19
COMPLEMENTING EXISTING REFERENCES.....	20
AN ILLUSTRATIVE EVALUATION DESIGN: THE MOTIV™ TKA CIRCLE.....	20
THE CIRCLE DATASETS.....	20
ILLUSTRATIVE EVALUATION.....	20
<i>The Question</i>	20
<i>Design</i>	21
PRE-SPECIFIED ANALYSIS	21
WHAT THE RESULT WOULD SHOW	21
CONDITIONS.....	22
WHY DEMAND SCALES WITH MODELS, VERSIONS AND DEPLOYMENTS	22
THE MULTIPLICATION.....	22
THE MARKET ALREADY PAYS FOR HUMAN JUDGMENT.....	23
HEALTH-DATA LICENSING IS ESTABLISHED	23
WHO BENEFITS	23
CONDITIONS FOR SOUND EVALUATION	24
CONCLUSION	25
REFERENCES	26

EXECUTIVE SUMMARY

Artificial intelligence now runs through nearly every layer of healthcare. It drafts clinical notes, codes claims and reviews prior authorizations. It flags findings on images and predicts deterioration. By OpenAI's own analysis, more than 230 million people a week ask ChatGPT health and wellness questions.¹

Almost none of it has been tested against what happens to patients. Of 1,357 AI-enabled medical devices cleared by the U.S. Food and Drug Administration (FDA) through December 5, 2025, only 3 evaluated patient-centered outcomes such as mortality, morbidity or readmissions.² The cause lies in the references available. Today every clinical AI system is trained on the record or graded against opinion:

- The record is claims, EHR entries and, increasingly, notes drafted by other models.
- Opinion means physician-written rubrics and expert labels. HealthBench, a benchmark for large language models in health, scores responses against rubrics written by 262 physicians and applied by a model-based grader.³

Neither reference measures what matters: whether the patient got better, stayed better and avoided harm. Expert opinion varies most in the ambiguous cases AI is deployed to triage. Records carry billing incentives and model errors, and rarely record outcomes at all. Public benchmarks leak into the training data of the models later tested on them. The developers of these systems agree. Researchers at Google and academic collaborators wrote in *Nature Medicine* in September 2026: "Trust in clinical artificial intelligence (AI) cannot be benchmarked into existence".⁴

A reference standard for clinical AI must be, at a minimum:

- Generated independently of any model's training data, and after it.
- Anchored in outcomes measured at defined intervals.
- Traceable to identified clinical observers.
- Drawn from multiple sites.
- Refreshed continuously, so that every new model version meets data it has not seen.

FDA's guidance on predetermined change control plans already calls for test data that are "never shown to the algorithm during training" and drawn from sites "different from those that were used to generate training and tuning data".⁵ The full set of ten requirements is developed in [What a Reference Standard for Clinical AI Must Be](#).

A Circle Dataset is a set of patient Cases captured prospectively at the point of care against a predefined Observational Protocol, with outcomes recorded at defined intervals. Circle Datasets meet the requirements above by construction:

Specified, Then Captured. Each Circle Dataset is built from an Observational Protocol defined before any data exists, and captured prospectively at the point of care.

Linked To Outcomes. Each Case links a patient’s diagnosis and treatment to outcomes over time.

Traceable. Every data point is entered by an identified clinician or patient and time-stamped against a specific question.

Uncontaminated. A Circle Dataset captured after a model’s training cutoff cannot have been ingested by that model.

Continuously Refreshed. Because Circles keep enrolling, each model version can be evaluated against Cases it has never seen.

These properties make Circle Datasets the reference for testing what matters most in clinical AI: prediction of outcomes, treatment selection, appropriateness of care, performance drift, and reasoning across a multi-visit patient journey. Demand for that reference scales with the number of models, versions and deployments, not with the number of patients. Health-system certification, state law, enforcement actions that require accuracy claims to be substantiated, litigation that puts model performance in discovery, and insurers that must price AI risk each reinforce it.

This paper sets out the evaluation gap, the requirements for a clinical AI reference standard, the parties who require one, and how Circle Datasets meet those requirements. It also states what Circle Datasets cannot test and the conditions their use requires. A companion paper, [*Circle Datasets for Well-Correlated Long-Term Healthcare Outcomes Capture*](#), addresses the outcome evidence itself.⁶

HEALTHCARE AI IS ALREADY EVERYWHERE

AI in healthcare is no longer a pilot category. It operates at each point where data is created,

interpreted, paid for and acted on.

Layer	Where AI Operates	Scale (latest available)
Regulated devices	Imaging, cardiology, neurology and other diagnostic and monitoring software	Over 1,600 AI-enabled devices on FDA's list as of September 2026, most through the 510(k) pathway. ^{2,7} FDA states that the list is not comprehensive. ⁸
Clinical documentation	Ambient scribes drafting notes from the clinical conversation	All 43 surveyed U.S. health systems had at least begun developing or piloting ambient documentation. ⁹ By June 2025, 62.6% of 2,784 Epic hospitals had adopted an ambient tool. ¹⁰ 29% of surveyed physicians reported using voice-based documentation or AI scribes. ¹¹
Physician practice	Documentation, summarization, decision support	Over 80% of surveyed physicians use AI in a professional context. ¹²
Revenue cycle	Coding, documentation review, denial management	63% of 101 organizations polled by HFMA in late 2024 use AI and automation in the revenue cycle. ¹³
Utilization review	Prior authorization and medical-necessity review	Medicare's WISeR model has tested technology-assisted prior authorization in six states since January 1, 2026. Licensed clinicians make final adverse determinations. ^{14, 15}
Consumer health	Conversational AI connected to personal records	More than 230 million people ask ChatGPT health and wellness questions each week, by OpenAI's analysis. ChatGPT Health launched in January 2026. ¹ OpenAI for Healthcare followed a day later, for health systems. ¹⁶
Data platforms	Foundation models and analytics trained on EHR and claims data	Epic's Curiosity model was trained on more than 100 billion patient medical events from Cosmos, a database of more than 300 million patients. ^{17,18} Tempus reported \$316.4 million of 2025 revenue from its data and applications segment. ¹⁹ NCQA is developing HEDIS measures on Komodo Health's AI platform. ²⁰

How The Layers Connect

The First Mile Is Now Machine-Written

Ambient documentation matters beyond its own risk profile because it sits at the start of the data chain. The note it drafts becomes the record that coding, analytics, registries and later models consume. When a model writes the first mile, every

downstream model is partly trained and evaluated on another model's output.

Decisions Are Increasingly Automated End To End

In prior authorization, AI can sit on both sides of a single decision: a provider-side tool assembles documentation, a payer-side tool scores it, and the dispute that follows is decided on records that neither side's model generated independently. The same structure applies to coding and denials, quality measurement and clinical decision support.

The Question Every Layer Now Faces

Each of these systems makes a claim about patients: this finding is present, this note is accurate, this service is appropriate, this patient is at risk, this treatment will help. Every such claim can be checked only against what happened to the patient. The rest of this paper concerns that verification, and its absence.

THE EVALUATION GAP

Cleared Devices Are Rarely Tested On Outcomes

A 2026 analysis of 1,357 AI-enabled devices cleared by FDA through December 5, 2025 found that “only 3 (0.2%) evaluated patient-centered outcomes such as mortality, morbidity, or readmissions”.² An earlier study of 521 FDA-authorized AI devices found that 226 (43%) lacked published clinical validation data.²¹

Most AI devices reach the market by showing substantial equivalence to a predicate device. Their evidence typically rests on retrospective accuracy benchmarks and surrogate endpoints, not on the effect on patients after deployment.²

Early Recalls Are Associated With Missing Validation

Recalls. Among 950 AI-enabled devices, 60 (6.3%) were associated with 182 recall events through November 2024. Of those events, 79 (43.4%) occurred within 12 months of clearance.²²

Validation. Lack of reported clinical validation was independently associated with recall (adjusted odds ratio 2.8, 95% CI 1.6–4.7).²²

Pre-market evidence should catch such failures, and post-market monitoring should correct them. Neither can do so without an independent reference.

Models Degrade When They Leave The Development Setting

The external validation of a widely implemented proprietary sepsis model is the standard example. Across 38,455 hospitalizations at Michigan Medicine, the model achieved an area under the curve of 0.63, below the performance reported by its developer. It missed 67% of patients with sepsis and generated alerts on 18% of all hospitalizations.²³ The model was in wide use before an independent evaluation measured it against the outcome it was meant to predict.

The Label Defines What A Model Learns

A widely used population-health algorithm assigned risk using predicted health care costs as its label. The authors found that “the bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients”.²⁴ The model fit its label and missed the patient’s health. The choice of reference standard is therefore not a technical detail. It determines what the system is optimized to do.

Generative AI Adds Errors At The Source

Production use. In a prospective pilot of an ambient AI scribe in routine outpatient care, physicians reviewed 356 notes and found accidental omissions in 18%, hallucinations in 11.5% and errors rated as posing serious or imminent risk in 5.3%. Of 960 notes for which the vendor’s editing metric was available, 14.9% were left entirely unedited.²⁵

Simulated encounters. Across five commercial scribe platforms tested on 14 simulated ambulatory encounters, the mean percent error in clinical notes was 26.3%.²⁶

These errors enter the record that later models learn from and are evaluated against.

Summary

The table summarizes the six symptoms.

Symptom	Evidence
Devices cleared without outcome testing	3 of 1,357 evaluated patient-centered outcomes. ²
Devices without published clinical validation	226 of 521 (43%). ²¹
Early recalls	43.4% of recall events within 12 months of clearance. ²²
Performance loss outside development	Sepsis model AUC 0.63; 67% of sepsis cases missed. ²³
Label bias	Health care cost used as a proxy for illness. ²⁴
Errors at the source	5.3% of reviewed AI-drafted notes with serious-risk errors. ²⁵

WHY CURRENT REFERENCES FAIL

No Model Can Be Verified Against Its Own Inputs

Every algorithm is a lossy, directional transformation. It selects, imputes, normalizes, discards and infers, and the input cannot be reconstructed from the output. This holds for a deterministic imaging algorithm as much as for a foundation model; the difference is one of degree, not kind.

It follows that no algorithmic output can be verified against the data that produced it. Verification requires a reference generated independently of the pipeline. The references in common use each fall short of that requirement in a specific way.

Public Benchmarks Are Contaminated And Saturated

Saturation. Stanford’s MedHELM evaluation, published in *Nature Medicine* in 2026, found that “[w]hile large language models (LLMs) achieve near-perfect scores on medical licensing exams, these evaluations inadequately reflect the complexity and diversity of real-world clinical practice”. ²⁷

Pattern-matching. When the correct answer to 68 medical questions was replaced with “None of the other answers,” all six models tested showed decreased accuracy, with

drops of 8.8% to 38.2%.²⁸ That pattern suggests reliance on pattern-matching over reasoning.

Contamination. Benchmark contamination, in which models “inadvertently incorporate evaluation benchmark information from their training data,” produces “inaccurate or unreliable performance during the evaluation phase”.²⁹

A benchmark published on the internet can become training data.

Expert Opinion Is Most Variable Where AI Is Used Most

Expert labels are a common reference in clinical AI. They are least reliable in the intermediate, ambiguous cases that AI is deployed to triage.

Domain	Finding
Breast pathology	Pathologists’ interpretations matched the reference diagnosis 96% of the time for invasive carcinoma and 48% for atypia. ³⁰
Melanocytic lesions	Accuracy against an expert-consensus reference was 92% for benign nevi and mildly atypical lesions and 25% for moderately dysplastic nevi. ³¹
Electrocardiography	Inter-reader kappa was 0.33 and sensitivity for STEMI against angiography was 65%, across 124 physicians. ³²
Chest radiography	Consensus among three radiologists rose from 41.8% after the initial read to 96.8% after up to five rounds of adjudication. ³³
Diagnosis in general	Median major-error rate of 23.5% in autopsy series from 1966 to 2002. ³⁴

Rubrics are opinion too. HealthBench scores responses against conversation-specific rubrics “created by 262 physicians,” applied by “a model-based grader, validated against physician judgment”.³ A rubric encodes what physicians believe a good answer contains. It does not record what happened to a patient who received that answer.

Adjudication is consensus, not truth. Adjudication raises agreement, as the chest-radiograph study shows, but agreement is not outcome. A label that three experts accept after five rounds of discussion is a consensus.

The Record Reflects Billing, Copying and Other Models

FDA notes that “[t]he purpose of medical claims data is to support payment for care”, and

that outcomes “representing mild symptoms or events occurring outside of medical care (e.g., out-of-hospital death) will not generally be well-captured”.³⁵ EHR notes are largely copied, templated or imported. In one analysis of inpatient progress notes, 18% of text was manually entered.³⁶ In a decade of outpatient notes at one academic center, about 29% of text in 2018 was typed directly.³⁷ A growing share is now drafted by ambient AI, as described in Healthcare AI Is Already Everywhere.

Training and evaluating models on such records compounds the problem. Models trained on indiscriminately generated model output develop defects in which “tails of the original content distribution disappear”.³⁸ In medicine those tails are the rare presentations, atypical responses and adverse events that matter most.

Synthetic Data Inherits Its Source

Synthetic patient records are useful for software testing and privacy-preserving development. As a reference standard they fail for a structural reason: they are generated by a model from existing data. They reproduce that data’s omissions and biases, and they cannot contain an outcome that the source data did not record.

What The Failures Have In Common

The table compares each reference with three properties an evaluation reference needs.

Reference	Independent of the model?	Post-dates training?	Anchored in outcomes?
Public benchmark	Partly	No; absorbed over time	No
Expert labels and rubrics	Yes	Sometimes	No; records opinion
EHR and claims	No; often the training source	No	Rarely
Synthetic data	No	No	No
Prospective, protocol-defined outcomes	Yes	Yes	Yes

WHAT A REFERENCE STANDARD FOR CLINICAL AI MUST BE

Regulators and standards bodies already describe what an evaluation reference must do. Taken together, their language defines ten requirements.

Ten Requirements

#	Requirement	Basis
1	Independent of training data: never shown to the model	FDA PCCP guidance. ⁵
2	Post-dates the model's training cutoff, so it cannot have leaked	Contamination evidence. ²⁹
3	Anchored in outcomes: the true condition, measured over time, not opinion alone	FDA lifecycle draft, "true condition"; ³⁹ label-bias evidence. ²⁴
4	Preserves reference-standard uncertainty: individual observations, not only consensus	FDA lifecycle draft. ³⁹
5	Multi-site, from sites other than training sites	FDA PCCP guidance. ⁵
6	Representative and complete for the intended use	EU AI Act, Article 10. ⁴¹
7	Longitudinal, so drift and delayed outcomes can be measured	FDA lifecycle draft. ³⁹
8	Traceable to identified observers, with source-data integrity	FDA and ICH source-data standards. ^{44, 45}
9	Refreshed, so each model version meets data it has not seen	FDA PCCP version comparison. ⁵
10	Licensed with clear rights and consent for evaluation use	Privacy and consent law; see Conditions for Sound Evaluation.

No benchmark, expert-label set, EHR extract or synthetic dataset meets all ten. Each falls short on at least one of requirements 2, 3 and 9, which matter most for models that are updated continuously and deployed at scale.

What The Frameworks Say

FDA, predetermined change control plans (issued December 2024, reissued August 2025)

- Test data "are never shown to the algorithm during training".

- Test data “should be independent of data used for training and tuning”.
- Test data “should generally be from multiple sites different from those that were used to generate training and tuning data”.
- A modified device is tested against both the original device and “the last modified version”.⁵

FDA, AI-enabled Device Software Lifecycle Draft Guidance

The guidance addresses performance “throughout the TPLC, including in the postmarket setting” (the total product life cycle). It notes that devices “can be sensitive to differences in input data (also referred to as data drift)”. It defines a reference standard as “the best available representative truth that can be used to define the true condition for each patient/case/record,” and asks sponsors to describe clinician grading and adjudication, intra- and inter-clinician variability, and “the uncertainty inherent in the selected reference standard”.³⁹

Coalition for Health AI, Testing and Evaluation Framework For Clinical Decision Support

For clinical decision support the coalition lists “Evidence Concordance, Citation Accuracy, and Guideline Alignment,” with a suggested benchmark of “≥ 90% evidence concordance with specialty guidelines”.⁴⁰

EU AI Act, Article 10(3)

“Training, validation and testing data sets shall be relevant, sufficiently representative, and to the best extent possible, free of errors and complete in view of the intended purpose”.⁴¹

ONC HTI-1

Under 45 CFR 170.315(b)(11), certified health IT that supplies predictive decision support must expose source attributes, including “[v]alidity of intervention in data external to or from a different source than the initial training data,” and must describe how validity and fairness are monitored over time.⁴² A December 2025 proposed rule (HTI-5) would “fully remove the artificial intelligence (AI) ‘model card’ requirements” from that criterion.⁴³ The underlying need for external validation remains.

WHO REQUIRES OUTCOME-ANCHORED EVALUATION

Demand comes from six directions. Some are mandates; more are accountability

mechanisms that reward evidence without requiring it. As formal federal requirements loosen in some areas, the other channels carry more of the load.

Regulators

FDA

- The predetermined change control plan framework ties each update made under an authorized plan to testing on independent data.⁵
- In September 2025, FDA requested public comment on measuring real-world performance of AI-enabled devices, asking among other things which data sources are used for ongoing performance evaluation and what triggers additional assessment.⁴⁶
- In August 2026, FDA opened a public comment request on generative-AI-enabled devices. Its announcement describes a “competency assessment,” “non-clinical device benchmarking and clinical confirmation,” and “risk-proportionate postmarket monitoring”.⁴⁷
- In 2025, predetermined change control plans appeared in about 10% of AI 510(k) clearances (30 of 295 through December 20, by one industry tally).⁴⁸ Each plan commits the manufacturer to testing every future update, and each test needs a reference.

European Union

The AI Act’s data-quality obligations in Article 10 apply to high-risk AI systems, which include many AI medical devices.⁴¹ The Digital Omnibus on AI, in force since July 27, 2026, set the application date for high-risk systems in products covered by Annex I of the Act, a list that includes medical devices, at August 2, 2028, and for Annex III systems at December 2, 2027.⁴⁹

Health Systems

Guidance. Joint Commission and CHAI guidance (September 2025) calls on hospitals to request from vendors information on “how AI tools were tested and validated for their intended use,” and to monitor locally, including “[r]egularly validating and testing AI tools for relevant performance and reliability”.⁵⁰

Certification. The Joint Commission’s voluntary Responsible Use of AI in Healthcare certification (June 2026) includes “monitoring, evaluating, and validating safety performance, effectiveness, and responsible use”.⁵¹

Liability. A health system that deploys AI carries the clinical and legal consequences of its errors, and it needs evidence about the model’s performance in its own patients.

Payers and States

California. SB 1120 requires that a health plan’s utilization-review AI tool be “open to inspection for audit or compliance reviews by the department” and that its “performance, use, and outcomes are periodically reviewed and revised to maximize accuracy and reliability”.⁵²

Utah. HB 452 (effective May 7, 2025) gives suppliers of mental-health chatbots an affirmative defense in licensing actions if, among other conditions, they maintain and file a policy that provides for testing to ensure the chatbot’s output “poses no greater risk to a user than that posed to an individual in therapy with a licensed mental health therapist,” and that “reasonably ensures regular, objective reviews of safety, accuracy, and efficacy”.⁵³

Texas. The Texas Responsible Artificial Intelligence Governance Act (TRAIGA), effective January 1, 2026, lets the Attorney General request, after a complaint, “any metrics the person uses to evaluate the performance of the artificial intelligence system” and a description of its post-deployment monitoring.⁵⁴ SB 1188, effective September 1, 2025, conditions a practitioner’s use of AI for diagnostic purposes on the practitioner’s review of “all records created with artificial intelligence” in a manner consistent with Texas Medical Board standards.⁵⁵

Uncertainty. State action is not uniform, and federal policy is moving against some of it. Colorado replaced its AI Act with a narrower automated-decision-making law, SB26-189, signed May 14, 2026 and effective January 1, 2027.⁵⁶ Executive Order 14365, signed December 11, 2025, directs the Department of Justice to establish a task force to challenge state AI laws and the Department of Commerce to identify “onerous” ones.⁵⁷ Statutes may narrow or fall. However, the certification, enforcement, litigation and insurance channels below do not depend on them.

Enforcement and Litigation

Substantiating accuracy claims

- Texas’s 2024 settlement with Pieces Technologies, a healthcare generative-AI company, requires that when its marketing cites metrics describing its products’ outputs, it disclose “the meaning or definition of such metric, benchmark, or similar

measurement” and “the method, procedure, or any other process used ... to calculate the metric”.^{58, 59}

- The FTC’s final order against Workado requires the company to hold “competent and reliable evidence” for claims about the accuracy of its AI content detector.⁶⁰ The FTC’s action against IntelliVision alleged that the company “did not have support for its claim that its software was free of gender or racial bias”.⁶¹
- The court-approved tenant-screening settlement with SafeRent (November 2024) requires that any new screening score the company wishes to use be validated by a third party that the plaintiffs agree to, as summarized by class counsel.^{62, 63}

Litigation

- In litigation over an insurer’s post-acute care algorithm, a federal magistrate judge in March 2026 compelled production of documents on the model’s development, implementation and oversight, but not its source code or underlying data.⁶⁴
- Investors filed 15 AI-related securities class actions in the first half of 2026, which Cornerstone Research describes as “on pace to nearly double the 2025 total”.⁶⁵
- Disputes of this kind turn on evidence of how a model performed. Evidence that is independent of the defendant and anchored in outcomes answers that question most directly.

Insurers

Liability insurers are beginning to exclude generative-AI exposures from standard forms: in January 2026, optional ISO general-liability endorsements became available that exclude them.⁶⁶ Specialist carriers offer affirmative cover instead. A Chaucer and Armilla AI product announced in April 2025 covers AI failures such as “hallucinations” and “model drift,” and the announcement points to Armilla’s “tools for evaluating and benchmarking AI model performance”.⁶⁷ An insurer that must price AI error needs the same independent measurement of model performance that a regulator or a court would ask for.

AI Developers

Google. Researchers at Google and academic collaborators wrote in *Nature Medicine* that “[t]rust in clinical artificial intelligence (AI) cannot be benchmarked into existence”.⁴ A Google-led prospective feasibility study of the conversational diagnostic system AMIE had 100 patients complete AMIE interactions in an ambulatory primary care clinic (preprint).⁶⁸

Microsoft. Microsoft AI states that its diagnostic orchestrator “would only be approved after rigorous safety testing, clinical validation, and regulatory reviews”.⁶⁹

Stanford. MedHELM, a Stanford-led evaluation, concludes that exam-style evaluation does not reflect clinical practice.²⁷

Each points to the same need: validation in real clinical settings.

CIRCLE DATASETS AS EVALUATION DATASETS

How Circle Datasets Are Produced

The Circles Platform starts with the clinical question and structures capture to answer it.⁷⁰

Observational Protocol. A Protocol is a set of surveys made up of canonical question-and-answer pairs. It specifies what will be asked, of whom, by whom and when. Each pair carries inheritable Attributes: FHIR resources, medical codes, the treating clinician and any other pertinent element.

Circle. A Circle is the group of providers collecting clinical data as specified in a single Observational Protocol, and whose patients provide the long-term outcomes data correlated to those clinical question-and-answer pairs. A Circle Member may work in a small practice or an academic medical center, anywhere in the world.

Case. A Case is one patient’s complete journey against the Protocol: diagnosis, treatment, complications and outcomes at defined intervals.

Circle Dataset. A Circle Dataset is any set of Cases assembled by Attribute or question-and-answer pair, traceable to source. One Protocol yields many Circle Datasets without new collection.

Closed System. The patented Circles Platform is a closed system. Every data point is entered directly by an identified Circle Member (a clinician, a patient or a laboratory) and time-stamped against a specific question. That meets the attributable, contemporaneous and original standard applied to source data in regulated research.^{44, 45}

Sustained Contribution. Circle Members co-own the Circle Datasets their Cases produce and receive a substantial portion of net licensing revenue, allocated in proportion to the volume and completeness of their contributions. Patients see their own recovery

against a relevant cohort. Both have reason to keep contributing over the months and years that outcomes require.

Properties That Meet The Requirements

The table maps each of the ten requirements to the Circle Dataset property that meets it.

Requirement	Circle Dataset Property
1. Independent of training data	Captured in a closed system against a Protocol, not scraped or brokered; licensed under access controls
2. Post-dates the training cutoff	Captured prospectively; Cases enrolled after a model's cutoff cannot have been ingested by it
3. Anchored in outcomes	Each Case links diagnosis and treatment to outcomes at defined intervals on validated instruments
4. Preserves uncertainty	Every entry attributed to an identified observer; variation between clinicians can be measured, not hidden
5. Multi-site	Circles span practices, institutions and settings
6. Representative and complete	Protocols define eligibility; completion is reported at each interval
7. Longitudinal	Follow-up at 6 weeks through 36 months, as the Protocol specifies
8. Traceable	Source-level provenance for every data point
9. Refreshed	Continuous enrollment supplies new Cases for each model version
10. Rights and consent	Consent terms travel with the data; Circle Members co-own the datasets

Three Properties No Other Reference Combines

A Perpetual Holdout. A Circle Dataset captured after a model's training cutoff cannot have leaked into it, and new Cases keep arriving. Each model version can therefore be tested on data that did not exist when it was trained. This is the condition FDA's change-control framework asks for,⁵ met continuously rather than once.

Outcomes, Not Opinion. Where a benchmark asks whether an answer matches what physicians would say, a Circle Case can ask whether the prediction or recommendation matched what happened to the patient. This is the "true condition"

standard in FDA’s draft lifecycle guidance,³⁹ applied to the questions clinical AI is increasingly asked to answer: who will improve, which treatment will work, and whether care was appropriate.

Disagreement Preserved. Attribution records who observed what. Evaluators can compare a model with individual clinicians, measure how much clinicians vary on the same kind of Case, and state the uncertainty of the reference itself, as FDA’s draft lifecycle guidance asks.³⁹ A consensus label conceals that variation. An attributed record exposes it.

Delivery Options

Licensed holdout sets. A licensee receives a sequestered dataset for a defined evaluation, with no training use.

Evaluation as a service. A model is run against Circle Datasets, and results are returned without the underlying records.

Federated evaluation. Computation runs where the records sit, inside a local data node, and only verified results leave. This extends evaluation to institutions and jurisdictions that do not release patient-level data.

WHAT CIRCLE DATASETS CAN AND CANNOT TEST

An evaluation dataset is valuable to the degree that it matches the task. Circle Datasets are structured, longitudinal records of diagnosis, treatment and outcome. They are strongest wherever a model makes a claim about what will happen, or should happen, to a patient.

Tasks Which Circle Datasets Are Built To Evaluate

Task	What Is Tested	Example
Outcome prediction and prognosis	Given baseline information, does the model predict the outcome observed?	Predicted versus observed function and pain six months after knee replacement
Treatment selection	Does the model’s recommendation correspond to the treatment associated with better outcomes in comparable patients?	Recommended pain-management pathway versus outcomes observed across pathways
Appropriateness and prior authorization	Does the model’s determination of necessity align with documented indications and subsequent outcomes?	Approval or denial against pre-treatment findings and post-treatment results

Task	What Is Tested	Example
Risk stratification	Are patients flagged as high-risk those who later experience complications?	Complication risk scores against recorded complications
Drift monitoring	Does performance change as practice, populations or products change?	Model accuracy on successive enrollment cohorts
Longitudinal reasoning by agents	Can an agent reason across multiple visits and revise its assessment as new information arrives?	Staged presentation of a Case, visit by visit, scored against the eventual outcome
Documentation fidelity	Do AI-drafted notes and extracted fields agree with structured entries made by the clinician?	Scribe-extracted findings versus Protocol answers for the same encounter (where both exist)
Subgroup performance	Does performance hold across age, sex, comorbidity, site and setting?	Stratified accuracy by Case Attributes

How A Case Becomes An Evaluation Item

The longitudinal structure of a Case allows it to be presented to a model in stages:

Baseline presentation. The model receives the information available at the first visit: history, examination, baseline scores and diagnosis.

Model output. The model produces a prediction, a recommendation or a determination.

Staged updates. Subsequent visits are revealed one at a time, and the model may revise its output.

Scoring against outcome. Outputs are scored against the outcomes recorded at the Observational Protocol's intervals, and against the recorded decisions of the clinicians who treated the patient.

The same Case can therefore test a single prediction, a sequence of decisions, or a model's stability as evidence accumulates.

Tasks Which Circle Datasets Do Not Address

Circle Cases do not contain images, waveforms, audio or conversation transcripts, and they are not a corpus of clinical free text. They are not suited to evaluating image segmentation or detection, speech recognition, conversational fluency or retrieval over document

collections. For systems that do those things, a Circle Dataset can evaluate the clinical claim the system ultimately makes, such as a diagnosis, a risk estimate or a recommendation, but not the intermediate perception or language steps.

Complementing Existing References

Outcome-anchored evaluation does not replace expert review or benchmark testing. Expert review establishes whether an answer is reasonable. Benchmarks establish baseline competence. Outcome data establishes whether the model's claims about patients are borne out. Clinical AI needs all three, and the third is the one now missing.

AN ILLUSTRATIVE EVALUATION DESIGN: THE MOTIV™ TKA CIRCLE

This section describes how an evaluation could be run on an active set of Circle Datasets. No results are reported.

The Circle Datasets

The [MOTIV™](#) total knee arthroplasty (TKA) Circle is a national program overseen by RegenMed and the [Orthopaedic Research and Education Foundation](#) (OREF), launched in early 2026 and described in the *Journal of Orthopaedic Experience and Innovation*.⁷¹

Protocol. Practicing surgeons authored the Observational Protocol and OREF's experts reviewed it. It records baseline status, surgical and perioperative details, complications, and patient-reported outcomes at defined intervals.

Scale. As of September 2026, 305 patients were enrolled across more than 15 surgeon Circle Members.

Regulatory basis. A WCG IRB determination under 45 CFR 46.104(d)(4), issued February 3, 2026, covers multiple Circle Member sites.⁷²

Follow-up. Patient-reported outcome completion has reached 85% at six weeks, 89% at three months and 73% at six months.

Illustrative Evaluation

The Question

Do leading clinical AI systems predict six-month recovery after knee replacement from pre-operative information, and does their physician-rubric score predict their outcome accuracy?

Design

Holdout. Cases with six-month outcomes, enrolled after the training cutoffs of the models tested, are sequestered. No model has seen them.

Presentation. Each model receives the pre-operative portion of each Case as a structured clinical summary: demographics, comorbidities, baseline function and pain scores, imaging findings where recorded, and the planned procedure.

Outputs. Each model predicts the six-month function and pain scores, estimates the probability that the patient reaches a clinically meaningful improvement, and states a rationale.

Rubric scoring. Practicing orthopedic surgeons write rubrics for a subset of Cases, as a physician-rubric benchmark would. The model outputs are graded against them.

Outcome scoring. The same outputs are scored against observed six-month outcomes for error in predicted scores, and for discrimination and calibration on clinically meaningful improvement.

Clinician comparison. Where the Protocol captures the surgeon's own pre-operative expectation, model performance is compared with clinician performance on the same Cases.

Staged extension. Six-week and three-month data are revealed in sequence to test whether models revise their predictions appropriately.

Subgroups. Performance is reported by site, age, sex, body-mass category and baseline severity.

Pre-specified Analysis

Primary endpoint. The correlation between rubric score and outcome accuracy across models and Cases.

Secondary endpoints. Calibration, discrimination, subgroup gaps, and change in accuracy with staged information.

The analysis plan is fixed before any model output is scored.

What The Result Would Show

If rubric scores and outcome accuracy diverge, the evaluation shows that current

benchmarks can rank models differently from how the models perform for patients. That tests this paper's central claim directly. If they converge, it establishes when physician-rubric evaluation is a sufficient proxy. Either result would inform developers, health systems and regulators. The design generalizes to any Circle whose Protocol records a baseline, a treatment and a defined outcome.

Conditions

The MOTIV™ exemption determination applies on the condition that data are not submitted to FDA.⁷² Twelve-month outcomes, available from early 2027, extend the design to longer horizons.

WHY DEMAND SCALES WITH MODELS, VERSIONS AND DEPLOYMENTS

The Multiplication

Demand for most clinical data grows with patient volume. Demand for evaluation data grows with the number of claims being made about patients. Each of the following needs its own evaluation:

Each model. Every device, decision-support tool, scribe, prior-authorization engine and consumer assistant makes different claims.

Each version. A predetermined change control plan compares every modification with the original device and the last modified version on independent test data.⁵

Each deployment. Health systems are expected to validate locally and monitor over time.^{50, 51}

Each subgroup and site. Performance that holds on average can fail for particular populations or settings.^{24, 41}

Each period. Drift is detected only by repeated measurement against fresh data.^{39, 46}

The product of these factors grows faster than the number of patients treated. Because evaluation data must post-date each model version, it is consumed as it is used. A holdout set that a model has been tested against cannot serve as an unseen test for that model again.

The Market Already Pays For Human Judgment

AI developers already buy expert human data at scale, for training and for judgment-based evaluation. The table gives the latest reported figures; most are press reports, not audited results.

Company	Reported Scale
Scale AI	Revenue just shy of \$1 billion in 2025; Meta took a 49% stake in a deal reported at \$14 billion. ⁷³
Mercor	Annualized revenue run rate above \$2 billion (gross; CEO statement); reported talks at a \$20 billion valuation. ⁷⁴
Handshake AI	Nearly \$1 billion in gross annualized AI-training revenue; net revenue approaching \$300 million (press-reported). ⁷⁵
Snorkel AI	\$375 million annualized revenue run rate; \$3.5 billion valuation (company statement, September 2026). ⁷⁶

Mercor’s listing for its Physician Talent Network offers \$110 to \$250 an hour to physicians who “[w]rite gold-standard clinical responses, realistic patient cases and evaluation rubrics”.⁷⁷ That is the price of physician opinion produced on demand. Outcome truth is scarcer. It cannot be produced by the hour, because it must be observed over months or years in real patients.

Health-Data Licensing Is Established

Health data is already licensed to AI and life-sciences developers. Tempus reported \$316.4 million of data and applications revenue in 2025.¹⁹ AstraZeneca and Pathos agreed to pay Tempus “\$200 million in data licensing and model development fees” to build an oncology foundation model.⁷⁸ These transactions license records assembled from care delivery. Circle Datasets add what those records lack: prospective, protocol-defined outcomes that post-date any model’s training.

Who Benefits

In the Circles model, the clinicians who create the data co-own the Circle Datasets and receive a substantial portion of net licensing revenue, in proportion to their contributions. As demand for evaluation grows, the clinicians and patients who record outcomes share in the value of the reference that clinical AI requires.

CONDITIONS FOR SOUND EVALUATION

Scope. Circle Datasets evaluate clinical claims about patients, not perception or language tasks. Each evaluation should therefore state the task it covers. Broad evaluation across specialties requires Circles in those specialties; the Circles Platform itself does not depend on specialty.

Precision. Evaluation datasets should be released only where Case counts support the statistical precision the evaluation requires, and that precision should be reported with each result.

Confounding. Outcomes reflect the treatments patients received in routine care and are subject to confounding and selection. An outcome is the truth about that patient, not a counterfactual about what another treatment would have produced. Tasks should be framed accordingly. Prediction tasks are scored against observed outcomes. Treatment-selection tasks use comparative methods with pre-specified adjustment for recorded confounders. Baseline characteristics are reported alongside results.

Missing Follow-Up. Not every patient completes every follow-up. MOTIV™'s 73% completion at six months leaves about a quarter unreported. Evaluation sets report completion, and analyses test whether results change when incomplete Cases are included under stated assumptions.

Consent and Regulatory Use. Evaluations intended to support regulatory submissions should usually be conducted under a non-exempt IRB pathway designed for that purpose. Evaluation by third parties must be covered by the consent under which the data were collected. Privacy litigation over ambient AI documentation shows the cost of capture without clear consent: a proposed class action filed in November 2025 alleges that Sharp HealthCare recorded exam-room conversations without patient consent.⁷⁹ The Circles Platform carries consent terms with the data, and Protocols specify permitted uses, including evaluation. Uses outside those terms require new consent.

Independence. In litigation and discovery, Circle Datasets are collected for clinical purposes by an independent platform, not by the model developer. That independence gives them weight, and their provenance is documented for that reason.

Compensation. Paying clinicians who contribute data may draw scrutiny under the Anti-Kickback Statute and the Stark Law. Circles compensation follows only the volume and completeness of contributed data and never depends on products, prescribing,

patient selection or referrals. Counsel should review any such arrangement.

CONCLUSION

Healthcare has adopted AI faster than it has built the means to check it. Models now draft the record, code the claim, review the authorization, flag the image, predict the deterioration and answer the patient. The evidence that these systems help patients is thin:

- 3 of 1,357 cleared devices evaluated patient-centered outcomes.
- An external validation of a widely implemented sepsis model fell short of its developer's reported performance.
- Benchmarks leak into the models they test.
- Rubrics measure agreement with physicians, not what happened to patients.

The missing element is a reference standard. It must be generated independently of any model and after it, anchored in outcomes, traceable to identified observers, drawn from many sites and refreshed continuously. Regulators describe it in their guidance. Health systems can be certified against it. States, enforcement agencies and courts increasingly ask for evidence of how models perform, and insurers need it to price risk. The developers of frontier clinical AI state that trust cannot be benchmarked into existence.

Circle Datasets provide that reference by construction:

- Physicians and patients are verifiably “in the loop” at the critical moments.
- Observational Protocols define the question before any data exists.
- Cases link diagnosis and treatment to outcomes over time.
- A closed system of identified contributors provides source-level provenance.
- Continuous enrollment supplies data that no model version has seen.
- The clinicians and patients who create the data share in its value, which sustains follow-up long enough to produce outcomes.

As the number of clinical models, versions and deployments grows, so does the need for a reference that no model produced. Outcome truth accumulates only with time. The organizations that begin capturing it now will be positioned to supply the reference against which healthcare AI is judged.

REFERENCES

1. OpenAI. [Introducing ChatGPT Health](#). January 7, 2026.
2. Abulibdeh R, Cajas Ordóñez SA, Celi LA, et al. [1,357 AI medical devices cleared, 3 actually tested on patient outcomes](#). PLOS Digit Health. 2026. doi:10.1371/journal.pdig.0001597.
3. Arora RK, Wei J, Hicks RS, et al. [HealthBench: Evaluating Large Language Models Towards Improved Human Health](#). arXiv:2505.08775. May 2025 (preprint).
4. Schaekermann M, Palepu A, Rodman A, et al. [Prospective evidence for conversational medical AI is hard, but non-negotiable](#). Nat Med. Published online September 14, 2026. doi:10.1038/s41591-026-04639-5.
5. U.S. Food and Drug Administration. [Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions](#). Final guidance, originally issued December 4, 2024; reissued August 18, 2025.
6. RegenMed, Inc. Circle Datasets for Well-Correlated Long-Term Healthcare Outcomes Capture. White paper, September 2026.
7. U.S. Food and Drug Administration. [Artificial Intelligence-Enabled Medical Devices](#). Digital Health Center of Excellence. Content current as of September 22, 2026.
8. U.S. Food and Drug Administration. [Artificial Intelligence-Enabled Medical Devices](#). Device list, last updated September 4, 2026.
9. Poon EG, Lemak CH, Rojas JC, Guptill J, Classen D. [Adoption of artificial intelligence in healthcare: survey of health system priorities, successes, and challenges](#). J Am Med Inform Assoc. 2025;32(7):1093–1100. doi:10.1093/jamia/ocaf065.
10. Yang F, Graetz I. [Ambient AI tool adoption in US hospitals and associated factors](#). Am J Manag Care. 2026;32(1):e25–e30. doi:10.37765/ajmc.2026.89876.
11. Doximity. [2026 State of AI in Medicine](#). March 17, 2026.
12. American Medical Association. [2026 Physician Survey on Augmented Intelligence](#) (Physician AI Sentiment Report). March 2026.
13. Healthcare Financial Management Association. [Most healthcare organizations are adopting AI in the revenue cycle: HFMA poll](#). May 30, 2025.
14. Centers for Medicare & Medicaid Services. [Wasteful and Inappropriate Service Reduction \(WiSeR\) Model](#). CMS Innovation Center model page. Accessed October 2026.
15. Centers for Medicare & Medicaid Services. [CMS launches new model to target wasteful, inappropriate services in Original Medicare](#). Press release, June 27, 2025.
16. OpenAI. [Introducing OpenAI for Healthcare](#). January 8, 2026.
17. Epic Systems. [Curiosity: A New Medical Intelligence for Clinical and Operational Insights](#). September 3, 2025.
18. Epic Systems. [Cosmos: About](#). Accessed October 2026.

19. Tempus AI. [Tempus Reports Fourth Quarter and Full Year 2025 Results](#). February 24, 2026.
20. Komodo Health. [Komodo Health and NCQA Collaborate to Accelerate the Future of Healthcare Quality Measurement](#). Press release, August 26, 2026.
21. Chouffani El Fassi S, Abdullah A, Fang Y, et al. [Not all AI health tools with regulatory authorization are clinically validated](#). Nat Med. 2024;30(10):2718–2720. doi:10.1038/s41591-024-03203-3.
22. Lee B, Kramer P, Sandri S, et al. [Early recalls and clinical validation gaps in artificial intelligence-enabled medical devices](#). JAMA Health Forum. 2025;6(8):e253172. doi:10.1001/jamahealthforum.2025.3172.
23. Wong A, Otles E, Donnelly JP, et al. [External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients](#). JAMA Intern Med. 2021;181(8):1065–1070. doi:10.1001/jamainternmed.2021.2626.
24. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. [Dissecting racial bias in an algorithm used to manage the health of populations](#). Science. 2019;366(6464):447–453. doi:10.1126/science.aax2342.
25. Taylor SL, Jost M, MacDonald S, et al. [Quality of clinical notes created by ambient listening generative AI: pragmatic prospective pilot study](#). JMIR Med Inform. 2026;14:e86474. doi:10.2196/86474.
26. Anderson TN, Mohan V, Dorr DA, et al. [Evaluating the quality and safety of ambient digital scribe platforms using simulated ambulatory encounters](#). Mayo Clin Proc Digit Health. 2025;3(4):100292. doi:10.1016/j.mcpdig.2025.100292.
27. Bedi S, Cui H, Fuentes M, et al. [Holistic evaluation of large language models for medical tasks with MedHELM](#). Nat Med. 2026;32(3):943–951. doi:10.1038/s41591-025-04151-2.
28. Bedi S, Jiang Y, Chung P, et al. [Fidelity of medical reasoning in large language models](#). JAMA Netw Open. 2025;8(8):e2526021. doi:10.1001/jamanetworkopen.2025.26021.
29. Xu C, Guan S, Greene D, Kechadi M-T. [Benchmark data contamination of large language models: a survey](#). arXiv:2406.04244. 2024 (preprint).
30. Elmore JG, Longton GM, Carney PA, et al. [Diagnostic concordance among pathologists interpreting breast biopsy specimens](#). JAMA. 2015;313(11):1122–1132. doi:10.1001/jama.2015.1405.
31. Elmore JG, Barnhill RL, Elder DE, et al. [Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study](#). BMJ. 2017;357:j2813. doi:10.1136/bmj.j2813.
32. McCabe JM, Armstrong EJ, Ku I, et al. [Physician accuracy in interpreting potential ST-segment elevation myocardial infarction electrocardiograms](#). J Am Heart Assoc. 2013;2(5):e000268. doi:10.1161/JAHA.113.000268.
33. Majkowska A, Mittal S, Steiner DF, et al. [Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation](#). Radiology. 2020;294(2):421–431. doi:10.1148/radiol.2019191293.

34. Shojania KG, Burton EC, McDonald KM, Goldman L. [Changes in rates of autopsy-detected diagnostic errors over time: a systematic review](#). JAMA. 2003;289(21):2849–2856. doi:10.1001/jama.289.21.2849.
35. U.S. Food and Drug Administration. [Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products](#). Final guidance, July 2024.
36. Wang MD, Khanna R, Najafi N. [Characterizing the source of text in electronic health record progress notes](#). JAMA Intern Med. 2017;177(8):1212–1213. doi:10.1001/jamainternmed.2017.1548.
37. Rule A, Bedrick S, Chiang MF, et al. [Length and redundancy of outpatient progress notes across a decade at an academic medical center](#). JAMA Netw Open. 2021;4(7):e2115334. doi:10.1001/jamanetworkopen.2021.15334.
38. Shumailov I, Shumaylov Z, Zhao Y, et al. [AI models collapse when trained on recursively generated data](#). Nature. 2024;631(8022):755–759. doi:10.1038/s41586-024-07566-y.
39. U.S. Food and Drug Administration. [Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations](#). Draft guidance, January 7, 2025.
40. Coalition for Health AI. [Testing and Evaluation \(T&E\) Framework: Clinical Decision Support](#). December 2025.
41. European Union. [Regulation \(EU\) 2024/1689 \(Artificial Intelligence Act\)](#), Article 10 and Annex I. Official Journal, July 12, 2024.
42. Electronic Code of Federal Regulations. [45 CFR 170.315\(b\)\(11\): Decision support interventions](#). Accessed October 2026.
43. Assistant Secretary for Technology Policy and Office of the National Coordinator for Health IT. [Health Data, Technology, and Interoperability: ASTP/ONC Deregulatory Actions To Unleash Prosperity](#). Proposed rule. 90 FR 60970, December 29, 2025.
44. U.S. Food and Drug Administration. [Electronic Source Data in Clinical Investigations](#). Guidance, September 2013.
45. International Council for Harmonisation. [E6\(R3\): Guideline for Good Clinical Practice](#). Step 4, January 6, 2025.
46. U.S. Food and Drug Administration. [Request for Public Comment: Measuring and Evaluating Artificial Intelligence-Enabled Medical Device Performance in the Real World](#). September 30, 2025. Docket FDA-2025-N-4203.
47. U.S. Food and Drug Administration. [FDA Seeks Public Feedback to Inform Regulatory Approach to Generative AI-Enabled Medical Devices](#). August 18, 2026. Docket FDA-2026-N-7874.
48. Innolitics. [2025 Year in Review: AI/ML Medical Device 510\(k\) Clearances](#). December 20, 2025 (industry analysis).

49. European Commission. [AI Omnibus enters into force](#). July 27, 2026. Regulation (EU) 2026/1744 of 8 July 2026.
50. The Joint Commission and Coalition for Health AI. [Guidance on the Responsible Use of AI in Healthcare](#). September 17, 2025.
51. The Joint Commission. [Responsible Use of AI in Healthcare certification](#). Press release, June 2026.
52. California Senate Bill 1120 (Ch. 879, Stats. 2024). [Health care coverage: utilization review](#). Health & Safety Code § 1367.01(k).
53. Utah House Bill 452 (2025). [Artificial Intelligence Amendments](#). Utah Code § 58-60-118. Enrolled text.
54. Texas House Bill 149, 89th Legislature (Texas Responsible Artificial Intelligence Governance Act). [Enrolled text](#). Tex. Bus. & Com. Code § 552.103.
55. Texas Senate Bill 1188, 89th Legislature. [Enrolled text](#). Tex. Health & Safety Code § 183.005.
56. Colorado General Assembly. [SB26-189: Automated Decision-Making Technology](#). 2026 Regular Session.
57. Executive Order 14365. [Ensuring a National Policy Framework for Artificial Intelligence](#). Signed December 11, 2025; 90 FR 58499, December 16, 2025.
58. State of Texas v. Pieces Technologies, Inc., Cause No. DC-24-13476 (Dallas County). [Assurance of Voluntary Compliance](#). Petition for approval filed August 21, 2024.
59. Office of the Texas Attorney General. [Attorney General Ken Paxton Reaches Settlement in First-of-its-Kind Healthcare Generative AI Investigation](#). September 18, 2024.
60. Federal Trade Commission. [FTC Approves Final Order against Workado, LLC, Which Misrepresented the Accuracy of its Artificial Intelligence Content Detection Product](#). August 28, 2025.
61. Federal Trade Commission. [FTC Takes Action Against IntelliVision Technologies for Deceptive Claims About its Facial Recognition Software](#). December 3, 2024.
62. Cohen Milstein. [Rental applicants using housing vouchers settle ground-breaking discrimination class action against SafeRent Solutions](#). Class counsel press release, April 2024.
63. Louis v. SafeRent Solutions, LLC, No. 1:22-cv-10800 (D. Mass.). [Order granting final approval of class settlement](#). November 20, 2024.
64. Estate of Lokken v. UnitedHealth Group, Inc., No. 0:23-cv-03514 (D. Minn.). [Order on plaintiffs' motion to compel discovery](#). Dkt. 162, March 9, 2026.
65. Cornerstone Research. [Securities Class Action Filings Surge in the First Half of 2026](#). July 29, 2026.
66. Independent Insurance Agents & Brokers of America. [Verisk to roll out new general liability exclusions for generative AI exposures](#). October 21, 2025.

67. Chaucer Group. [Chaucer co-develops new AI insurance product with Armilla AI](#). Press release, April 22, 2025.
68. Brodeur P, Koshy JM, Palepu A, et al. [A prospective clinical feasibility study of a conversational diagnostic AI in an ambulatory primary care clinic](#). arXiv:2603.08448. March 2026 (preprint).
69. Microsoft AI. [The Path to Medical Superintelligence](#). June 30, 2025.
70. RegenMed, Inc. Ground Truth for Human Health: An Overview of the Circles Platform. September 2026. [Detailed overview](#); [condensed version](#).
71. Mercuri J, Wickline A. [MOTIV™ and the next frontier of orthopaedic evidence generation: a new model for physician-led clinical research](#). J Orthop Exp Innov. 2026;7(2). doi:10.60118/001c.164991.
72. WCG IRB. Certificate of Action, MOTIV™ Project, February 3, 2026 (company record).
73. Nieva R. [Inside Scale AI's Business After Meta's Bombshell \\$14 Billion Deal](#). Forbes. May 14, 2026.
74. TechCrunch. [Mercor is in talks for a \\$20B valuation](#). July 9, 2026.
75. Dealroom. [Handshake's ARR crosses \\$1B as AI training revenue surges](#). April 14, 2026 (citing The Information).
76. TechCrunch. [Snorkel AI triples valuation to \\$3.5B as demand for AI training data booms](#). September 22, 2026.
77. Mercor. [Healthcare experts](#). Accessed October 2026.
78. Tempus AI. [Tempus Signs Expanded Strategic Agreements with AstraZeneca and Pathos](#). April 23, 2025.
79. KPBS. [Lawsuit claims Sharp HealthCare secretly recorded exam-room conversations without patient consent](#). December 11, 2025.