

MODEL MONITORING

After go-live: dataset shift, silent failure and the monitoring of clinical AI

ABSTRACT

A clinical model validated once is deployed into a population, a coding system and a data pipeline that keep changing. This review appraises the evidence on how fast and how visibly deployed models degrade. Drift arrives slowly, as calibration decays over years, and in step changes tied to identifiable events: a coding transition, a record-system change, a pandemic, a supplier's model update. Discrimination can hold while calibration fails, retrospective validation cannot see pipeline changes, and aggregate accuracy can conceal changed answers. The federal certification rule requires monitoring to be described, not to meet any standard. The paper sets out what a practice without a data team can actually watch, and what each signal misses.

Type Narrative review with operational implications **References** 29 **Reading time** 14 min

Last reviewed September 2026

1 Monitoring is asserted more often than it is specified

A clinical prediction model is validated on data from a period that has ended and then used on patients who have not yet arrived. Between the two, the population changes, practice patterns change, the codes that describe both change, and so does the software that carries data from the record to the model. Finlayson and colleagues supply the working definition: dataset shift occurs when a machine-learning system underperforms because of a mismatch between the data it was developed on and the data on which it is deployed.¹

Deployment documents usually answer this with one word: the model is “monitored”. Without a signal, a threshold, an owner and a response, the word describes an intention rather than a control. This paper asks how fast deployed clinical models degrade, how visibly, and what monitoring would have to consist of for a practice with no data-science team.

The short answer has four parts. Degradation is common and arrives at two speeds: slow calibration drift measured in years, and step changes that arrive abruptly when a code set, a record system, a pandemic or a supplier's model version changes. It is often quiet, because discrimination can hold steady while calibration fails. The federal certification rule requires a developer to describe how a model is monitored and sets no standard for what monitoring must find. And the monitoring available to a small practice is narrower than the word suggests, but not empty.

2 Three kinds of shift, and one that belongs to the pipeline

The terminology is inconsistent. Moreno-Torres and colleagues' unifying review distinguishes covariate shift, prior probability shift and concept shift.² In plain terms: under covariate shift the patients change while the relationship between their characteristics and the outcome holds; under prior probability shift, often called label shift, the outcome becomes more or less common; under concept shift the relationship itself changes, so the same patient would now carry a different correct answer. Feng and colleagues map monitoring onto the same three objects: the distribution of inputs, the distribution of outcomes, and the relationship between them.³

Finlayson and colleagues classify causes instead: changes in technology, in population and setting, and in behavior.¹ Many of their examples involve no change in the patients: high-sensitivity troponin assays, EHR updates that can alter internal variable definitions, the move from ICD-9 to ICD-10, and reimbursement that produced a measurable rise in documented sepsis. A fourth category is operationally distinct. Otles and colleagues call it infrastructure shift: changes in how data are accessed, extracted and transformed between the warehouse a model was validated on and the real-time feed it runs on.⁴ Nothing about the patient has moved; the pipe has.

TABLE 1

Types of shift, in plain terms, with a clinical example of each. The concept-shift row is illustrative; the others are documented in the cited studies.

Type	What changes	Example
Covariate shift	Who the patients are; the input–outcome relationship holds	Cancelled elective surgery in spring 2020 left a sicker inpatient mix, and sepsis alerts rose ⁵
Prior probability (label) shift	How common the outcome is	Changes in the rate of acute kidney injury tracked rising overprediction in Veterans Affairs models ⁶
Concept shift	What counts as the right answer for the same inputs	A model trained on past prescribing learned its training years' norms; the 2022 CDC opioid guideline updated the 2016 version and discourages rigid application of dosage thresholds ⁷
Measurement or infrastructure shift	How the same patient is recorded or delivered to the model	The ICD-10 transition; a 2008 record-keeping change; warehouse versus real-time extraction ^{4,8,9}

3 Drift arrives slowly, and then all at once

0.29

AUROC lost by a mortality model after a 2008 record-keeping change

1 in 6

diagnostic categories that shifted 20% or more at the ICD-10 transition, in two of three classification systems

+43%

sepsis alerts per day on a census that fell 35%

The slow version

Davis and colleagues developed seven models for hospital-acquired acute kidney injury on 2003 admissions to Veterans Affairs hospitals nationwide and validated each over the nine subsequent years.⁶ Discrimination was maintained for every model. Calibration declined as all of them increasingly overpredicted risk, most for the regression models, and changes in the rate of acute kidney injury were strongly linked to the overprediction. The authors draw a scheduling rule: update in response to periods of rapid drift rather than only at annual or biannual intervals.

The literature on keeping models working is thin. Guo and colleagues found 15 eligible studies among 4,457 publications on preserving performance under temporal dataset shift in clinical medicine.¹⁰ Calibration deteriorated more often than discrimination (11 studies against 3). Every mitigation strategy preserved calibration; effects on discrimination were inconsistent.

A figure to handle with care

A monitoring vendor's blog headline states that 91% of machine-learning models degrade in time, and the post presents the source as a study of how models behave after deployment.¹¹ The source is Vela and colleagues, who trained four standard model types on 32 datasets from four industries — healthcare operations, transportation, finance and weather — and observed temporal degradation in 91% of the 128 model–dataset pairs.¹² These were models trained experimentally on historical data, not deployed clinical systems. The study shows that degradation is the usual result in experiments of this kind. It does not estimate what share of deployed models degrade, which is how the headline reads.

The fast version

Step changes can be larger. Nestor and colleagues trained models on MIMIC-III intensive care records and tested them on later years.⁹ A system-wide change in record-keeping in 2008 cost a mortality model 0.29 in AUROC. Aggregating raw features into expert-defined clinical concepts cut the loss to 0.06: the model had learned the database's vocabulary as well as the patients' physiology.

Coding transitions do the same at national scale. Ellis and colleagues ran an interrupted time series on private insurance claims from 2010 to 2017, 2.1 billion enrollee person-months, across the October 2015 move from ICD-9-CM to ICD-10-CM.⁸ Level changes of 20% or more occurred in 20 of 127 HHS hierarchical condition categories (15.7%) and 46 of 282 AHRQ Clinical Classifications Software categories (16.3%), roughly 1 in 6, while the WHO disease chapters changed minimally. The monthly rate of people in the acute myocardial infarction category rose 131.5% (95% CI 124.1% to 138.8%), mainly because HHS added non-ST-elevation infarction diagnoses to it. The definition moved, and any model using that category as an input moved with it.

A pandemic did it in three weeks. Wong and colleagues compared alerts from the Epic Sepsis Model, a widely used proprietary model, across 24 hospitals in four health systems in the three weeks before and after each system's first COVID-19 case.⁵ Census fell 35%, from 10,159 to 6,634, while alerts per day rose 43%, from 953 to 1,363, counting at most one alert per patient per day; the share of patients alerting rose from 9% to 21%. The authors attribute the rise to cancelled elective surgery and higher acuity among those who remained. At the University of Michigan the model was deactivated in April 2020 because of spurious alerting.¹

Models learn the process of care

Models also learn how care is delivered. Agniel and colleagues examined 272 laboratory test types across 669,452 patients at two Boston hospitals.¹³ The presence of an order, regardless of its result, was significantly associated with survival for 233 of 272 tests (86%), and the timing of the order predicted survival more accurately than the result for 118 of 174 tests (68%). The supplement to the Finlayson letter compresses these into one claim, that timing mattered more than the value "in up to 86% of lab tests".¹ The 86% is the presence-of-order association; the timing-against-value comparison is 68%. Either way, a change in ordering habits is a change in a model's inputs even when no patient is different.

4 Why the failure is quiet

Five mechanisms keep degradation out of view, each defeating a different check.

Discrimination can hold while calibration fails

In the acute kidney injury models, discrimination held throughout the years in which calibration drifted.⁶ A validation report giving only an AUROC would have shown nothing wrong. Van Calster and colleagues argue that poorly calibrated algorithms can be misleading and potentially harmful for clinical decisions, and can be less useful than a well-calibrated competitor with a lower AUC.¹⁴ A risk score used at a threshold fails through calibration: the ranking survives, the number attached to each patient does not.

Retrospective validation cannot see the pipeline

Otles and colleagues applied a healthcare-associated infection model prospectively to 26,864 encounters from July 2020 to June 2021.⁴ AUROC was 0.767 (95% CI 0.737 to 0.801) against 0.778 (0.744 to 0.815) in the preceding year's retrospective validation, and the Brier score worsened from 0.163 to 0.189. The authors attribute the loss of discrimination mainly to infrastructure shift and the loss of calibration mainly to temporal shift, with confidence intervals on every component that include zero. The first attribution is the uncomfortable one: validation on warehouse data was partly validation on data the model would never see in production.

Performance monitoring is not drift detection

Kore and colleagues tested three drift-detection methods on real-world X-ray images, against natural drift from COVID-19 and synthetic drift.¹⁵ Monitoring performance alone was not a good proxy for detecting data drift, and detection depended heavily on sample size and patient features. The authors' case for watching inputs rests on settings where real-time evaluation is impractical, as when labels are costly.

Outcomes arrive late, and the model changes them

Feng and colleagues note that outcomes such as mortality need long follow-up, and that once clinicians act on a prediction it is hard to tell whether the model was wrong or the intervention prevented the outcome.³ Lenert, Matheny and Walsh state the consequence: the more effective a model and its intervention, the faster the model will appear to degrade, and refitting may not be the most effective response.¹⁶ A deployment that is working and one that is failing can draw the same declining curve.

Subgroups drift on their own schedule

Davis and colleagues tracked Veterans Affairs post-surgical models quarterly from 2014 to 2023, across 1,739,666 surgical cases, by self-reported race and sex.¹⁷ Fairness drift appeared in both original and updated models, and population-level updating sometimes narrowed a gap and sometimes widened it.

WHAT THE DEGRADATION LITERATURE DOES NOT ESTABLISH

Nearly every clinical model studied here was a hospital or intensive care model, and most studies are retrospective. None measured drift in an ambulatory specialty practice or in a practice of a few clinicians, where the numbers available to any method are small. The mechanisms transfer; the rates do not. No figure here should be read as the expected degradation of any particular product.

5 Language models change on someone else's calendar

A language model supplied as a service changes when its supplier does, as well as when its environment does. Chen, Zaharia and Zou evaluated the March 2023 and June 2023 versions of two widely used models on seven tasks, noting that when and how such models are updated is opaque.¹⁸ GPT-4 identified prime numbers with 84% accuracy in March and 51% in June. Their summary is that the behavior of the “same” service can change substantially in a relatively short time.

The medical-licensing results show the quieter form of the problem. In the preprint's full text, GPT-4's accuracy on licensing-examination questions fell from 86.6% to 82.4% and GPT-3.5 lost 0.8%, yet 12.2% of GPT-4's answers and 27.9% of GPT-3.5's differed between versions.¹⁸ An aggregate check would have registered almost nothing for GPT-3.5 while more than a quarter of its answers had changed. For a system whose outputs a clinician reads one at a time, the changed answers are the exposure.

Versions also end. One major supplier's documentation defines deprecation as the process of retiring a model, effective on announcement, and shutdown as the point at which it is no longer accessible.¹⁹ Its schedule shows a dated snapshot, `gpt-3.5-turbo-0613`, shut down on September 13, 2024 with the undated name `gpt-3.5-turbo` as its replacement, and an announcement on August 26, 2026 that the `whisper-1` transcription model will shut down on February 26, 2027. A drafting or transcription component can change with no decision by the practice: opaquely behind an undated name, or visibly at retirement.

The property that follows is modest: every generated output should carry the identifier of the exact model version that produced it, and a version change should be handled as a deployment event, with fixed local test cases run before and after. Without the identifier, a change in quality cannot even be dated. No study reviewed here validates edit rates or clinician acceptance as drift signals for generated text; they are plausible proxies, not tested ones.

6 What the rules and the guidance ask for

The certification and device regimes are set out in a companion paper; what matters here is what each says about monitoring.

Certification: describe it

The decision support interventions criterion of the ONC Health IT Certification Program requires a certified module to surface source attributes for predictive decision support. Under “Ongoing maintenance of intervention implementation and use”, they include a “Description of process and frequency by which the intervention's validity is monitored over time”, the “Validity of intervention in local data”, and the equivalent pair for fairness; a further heading requires a description of how often the

intervention is updated and its performance corrected when validity or fairness risks are identified.²⁰ These are disclosure duties. The criterion does not say what the process must detect, how often it must run, or what result would require action, and it binds the developer rather than the practice.

The eCFR records no amendment to the section since October 1, 2025, and its text still contains these attributes.²⁰ HTI-5, proposed on December 29, 2025, would reduce the criterion's scope to "fully remove the artificial intelligence (AI) 'model card' requirements"; comments closed February 27, 2026, and a search of the Federal Register in late September 2026 found no final rule.²¹ The monitoring disclosures are among the requirements at stake.

Device regulation: plan it

Where a software function is a device, FDA's Predetermined Change Control Plan guidance, issued December 4, 2024 and reissued August 18, 2025, lets planned modifications to an AI-enabled device software function be authorized in advance.²² A plan sets out the modifications, a protocol for data management, re-training, performance evaluation and updates, and an impact assessment; significant modifications outside it likely require a new submission. It governs changes a manufacturer decides to make. Drift the environment imposes is not a planned modification, and a function that is not a device needs no plan.

FDA's plainest statement on environmental drift is not guidance. Its September 30, 2025 request for public comment states that AI system performance "can be influenced by changes in clinical practice, patient demographics, data inputs, health care infrastructure, among other factors", and that retrospective testing and static benchmarks "are not designed to predict behavior in dynamic, real-world environments."²³ It says it is not draft or final guidance; comments closed December 1, 2025.

Voluntary guidance: scale it to risk

The most specific monitoring text in circulation is voluntary. The Joint Commission and the Coalition for Health AI released *The Responsible Use of AI in Healthcare* on September 17, 2025, with seven elements, one of them ongoing quality monitoring.²⁴ It asks for monitoring that is risk-based and scaled to the setting, checking tools that inform or drive clinical decisions more often than administrative or documentation tools. It lists regular validation, comparison of outputs against known parameters, checks that tools rely on up-to-date data, a dashboard and a route for reporting errors, and asks that monitoring responsibility be addressed in contracting, with vendors' model changes reported to whoever monitors. It calls itself initial and high-level. A voluntary Joint Commission certification under the same name launched in 2026.²⁵

What hospitals report

Practice trails even the voluntary text. Nong and colleagues analyzed the 2023 American Hospital Association information technology supplement: 65% of US hospitals used predictive models, 79% of those from their EHR developer.²⁶ Sixty-one percent of hospitals using models had evaluated them for

accuracy on their own data, and 44% for bias; hospitals that built their own models, had high operating margins or belonged to health systems were more likely to do so. These figures describe local evaluation for accuracy and bias, not recurring monitoring; this review found no comparable figure for physician practices.

7 What a practice without a data team can watch

The monitoring literature imagines hospital infrastructure; Feng and colleagues propose dedicated AI quality-improvement units.³ A practice of a few clinicians has a schedule, a record and its vendor's reports. Within those limits six signals are available, each detecting some kinds of shift and missing others (Table 2).

1. **Output volume per unit of activity.** Alerts, flags or drafts per 100 visits, per clinician, per week. The 2020 sepsis-alert study began from nurses' reports of overalerting and tested them by counting alerts.⁵ The signal needs no outcome data and moves within days.
2. **Input completeness and a calendar of known changes.** The rate at which each input is missing, and the dates on which inputs change by definition: code-set releases, system upgrades, new assays or templates. The next ICD-10-CM code set takes effect October 1, 2026.²⁷ The ICD-10 and MIMIC results show what such dates can do.^{8,9} Formal drift tests depend heavily on sample size, which a small practice lacks;¹⁵ a missingness rate and a calendar do not.
3. **Acceptance, override and edit rates.** Finlayson and colleagues ask clinicians to note misalignment between a model's predictions and their own judgment;¹ a rate turns that into a series. Override rates may also reflect interface design and attention, so a change is a prompt to look rather than a finding.
4. **Observed against expected, on the practice's own outcomes.** Where a model predicts something the practice records, mean predicted risk can be compared with the observed rate: calibration-in-the-large. A full calibration curve needs far more; Van Calster and colleagues cite a suggested minimum of 200 patients with the event and 200 without.¹⁴ Control charts such as CUSUM and EWMA, and dynamic calibration curves, turn repeated comparisons into alerts.^{3,28}
5. **Model and version identity.** The exact version behind every output, and the supplier's retirement schedule.¹⁹ It detects nothing by itself and makes every other signal interpretable.
6. **A fixed local test set.** A small set of de-identified or synthetic cases typical of the practice, with agreed answers, run at every version change and on a schedule. Youssef and colleagues argue that one-time external validation should give way to recurring local validation.²⁹

TABLE 2

Signals available without a data team: what each can detect, what it misses, and whether it needs outcome data.

Signal	Detects	Misses	Needs outcomes?
Output volume ⁵	Case-mix change; broken or altered feeds	Concept shift at a stable output rate	No
Input missingness and change calendar ^{8,27}	Measurement and infrastructure shift	Change in patients whose records are complete	No
Acceptance, override, edit rates ¹	Growing disagreement between model and clinicians	Errors the clinicians also miss	No
Observed against expected ¹⁴	Label shift; overall miscalibration	Miscalibration within risk bands, at small numbers	Yes
Version identity ¹⁹	Supplier-side model changes, by date	Everything else	No
Fixed local test set ^{18,29}	Changed behavior after a version change	Shift in the live population	Agreed answers only

WHAT THESE SIGNALS CANNOT DO

They detect change, not harm. A rise in alert volume may be the correct response to a sicker population, and a falling observed event rate may be the model working.¹⁶

Small numbers make every rate noisy, so tight thresholds will fire on chance. None of the six has been evaluated as a monitoring strategy for a small practice; they are what the evidence makes plausible, not what it has tested.

8 What “monitored” would have to mean

Stated as a property rather than a promise, monitoring for one model is a short specification: the signals, the baseline for each, the threshold that triggers review, the reviewer, and the actions available, at minimum to pause, recalibrate or retire. The Davis group supplies the schedule: update as calibration deteriorates rather than on a predetermined timetable.^{6,28} That means review on events as well as the calendar: a code-set release, a system upgrade, a supplier’s model change, a guideline revision, an abrupt change in who comes through the door.

Five statements are supported. Degradation is the usual finding in studies that look for it, at rates that do not transfer.^{6,10,12} The largest effects documented here came from step changes tied to identifiable events: a record-system change, a coding transition, a pandemic.^{5,8,9} Discrimination can hold while calibration

fails, so a stable AUROC is not evidence of a working model.^{6,14} Language-model services change on the supplier's schedule, and aggregate accuracy can conceal changed answers.¹⁸ And the certification rule requires a developer to describe monitoring, not to meet any standard in it.²⁰

What would settle the open questions is not another retrospective study of an inpatient model. It is prospective monitoring results from ambulatory deployments, suppliers disclosing version changes with before-and-after performance, and a threshold, anywhere in the rules, for what monitoring has to find. Until then "monitored" is a claim to be specified, in four parts: what is watched, against what baseline, how often, and who acts.

REFERENCES

Entries 4, 5, 6, 8, 9, 13 and 18 are the empirical studies that carry the argument about how fast and how visibly models degrade, and entry 10 is the only systematic review; entries 4 and 9 are peer-reviewed conference proceedings, and the medical-licensing figures in entry 18 are taken from its preprint version. Entries 20, 21, 22, 23 and 27 are a regulation, a proposed rule, agency guidance, a request for comment and a coding notice rather than evidence; entry 24 is voluntary guidance with no legal force, entry 25 is trade reporting used only for the 2026 certification launch, and entry 19 is a supplier's own documentation. Entry 12 is an experimental study outside clinical care and entry 11 is a vendor blog cited only to show how its figure is repeated; entries 3, 14, 16 and 29 are perspectives, and entry 7 is cited only as an example of a changed standard of care.

1. Finlayson SG, Subbaswamy A, Singh K, et al. *The Clinician and Dataset Shift in Artificial Intelligence*. New England Journal of Medicine. 2021;385(3):283–286. Letter, with online supplement. doi:10.1056/NEJMc2104626 POSITION PAPER
2. Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, et al. *A unifying view on dataset shift in classification*. Pattern Recognition. 2012;45(1):521–530. doi:10.1016/j.patcog.2011.06.019 REVIEW
3. Feng J, Phillips RV, Malenica I, et al. *Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare*. npj Digital Medicine. 2022;5:66. doi:10.1038/s41746-022-00611-y POSITION PAPER
4. Otles E, Oh J, Li B, et al. *Mind the Performance Gap: Examining Dataset Shift During Prospective Validation*. Proceedings of the 6th Machine Learning for Healthcare Conference. PMLR. 2021;149:506–534. proceedings.mlr.press COHORT
5. Wong A, Cao J, Lyons PG, et al. *Quantification of Sepsis Model Alerts in 24 US Hospitals Before and During the COVID-19 Pandemic*. JAMA Network Open. 2021;4(11):e2135286. doi:10.1001/jamanetworkopen.2021.35286 COHORT

6. Davis SE, Lasko TA, Chen G, et al. *Calibration drift in regression and machine learning models for acute kidney injury*. Journal of the American Medical Informatics Association. 2017;24(6):1052–1061. [doi:10.1093/jamia/ocx030](https://doi.org/10.1093/jamia/ocx030) RETROSPECTIVE COHORT
7. Dowell D, Ragan KR, Jones CM, et al. *CDC Clinical Practice Guideline for Prescribing Opioids for Pain — United States, 2022*. MMWR Recommendations and Reports. 2022;71(3):1–95. [doi:10.15585/mmwr.rr7103a1](https://doi.org/10.15585/mmwr.rr7103a1) GUIDELINE
8. Ellis RP, Hsu HE, Song C, et al. *Diagnostic Category Prevalence in 3 Classification Systems Across the Transition to the International Classification of Diseases, Tenth Revision, Clinical Modification*. JAMA Network Open. 2020;3(4):e202280. [doi:10.1001/jamanetworkopen.2020.2280](https://doi.org/10.1001/jamanetworkopen.2020.2280) QUASI-EXPERIMENTAL
9. Nestor B, McDermott MBA, Boag W, et al. *Feature Robustness in Non-stationary Health Records: Caveats to Deployable Model Performance in Common Clinical Machine Learning Tasks*. Proceedings of the 4th Machine Learning for Healthcare Conference. PMLR. 2019;106:381–405. proceedings.mlr.press RETROSPECTIVE COHORT
10. Guo LL, Pfohl SR, Fries J, et al. *Systematic Review of Approaches to Preserve Machine Learning Performance in the Presence of Temporal Dataset Shift in Clinical Medicine*. Applied Clinical Informatics. 2021;12(4):808–815. [doi:10.1055/s-0041-1735184](https://doi.org/10.1055/s-0041-1735184) SYSTEMATIC REVIEW
11. Viquez S. *91% of ML Models Degrade in Time*. NannyML blog. Apr. 11, 2023. nannyml.com INDUSTRY REPORT
12. Vela D, Sharp A, Zhang R, et al. *Temporal quality degradation in AI models*. Scientific Reports. 2022;12:11654. [doi:10.1038/s41598-022-15245-z](https://doi.org/10.1038/s41598-022-15245-z) BENCHMARK
13. Agniel D, Kohane IS, Weber GM. *Biases in electronic health record data due to processes within the healthcare system: retrospective observational study*. BMJ. 2018;361:k1479. [doi:10.1136/bmj.k1479](https://doi.org/10.1136/bmj.k1479) RETROSPECTIVE COHORT
14. Van Calster B, McLernon DJ, van Smeden M, et al. *Calibration: the Achilles heel of predictive analytics*. BMC Medicine. 2019;17:230. [doi:10.1186/s12916-019-1466-7](https://doi.org/10.1186/s12916-019-1466-7) POSITION PAPER
15. Kore A, Abbasi Babil E, Subasri V, et al. *Empirical data drift detection experiments on real-world medical imaging data*. Nature Communications. 2024;15:1887. [doi:10.1038/s41467-024-46142-w](https://doi.org/10.1038/s41467-024-46142-w) BENCHMARK
16. Lenert MC, Matheny ME, Walsh CG. *Prognostic models will be victims of their own success, unless....* Journal of the American Medical Informatics Association. 2019;26(12):1645–1650. [doi:10.1093/jamia/ocz145](https://doi.org/10.1093/jamia/ocz145) POSITION PAPER

17. Davis SE, Dorn C, Park DJ, Matheny ME. *Emerging algorithmic bias: fairness drift as the next dimension of model maintenance and sustainability*. Journal of the American Medical Informatics Association. 2025;32(5):845–854. [doi:10.1093/jamia/ocaf039](https://doi.org/10.1093/jamia/ocaf039) RETROSPECTIVE COHORT
18. Chen L, Zaharia M, Zou J. *How Is ChatGPT's Behavior Changing Over Time?* Harvard Data Science Review. 2024;6(2). [doi:10.1162/99608f92.5317da47](https://doi.org/10.1162/99608f92.5317da47). Medical-licensing figures from the preprint full text, §3.7, [arXiv:2307.09009v3](https://arxiv.org/abs/2307.09009v3). BENCHMARK
19. OpenAI. *Deprecations*. OpenAI API documentation; announcements of Nov. 6, 2023 and Aug. 26, 2026; accessed Sept. 2026. developers.openai.com INDUSTRY STATEMENT
20. Assistant Secretary for Technology Policy / Office of the National Coordinator for Health Information Technology. *Decision support interventions*. 45 C.F.R. § 170.315(b)(11)(iv)(B); eCFR, no amendment to § 170.315 recorded after Oct. 1, 2025 (checked Sept. 2026). ecfr.gov REGULATION
21. Assistant Secretary for Technology Policy / Office of the National Coordinator for Health Information Technology. *Health Data, Technology, and Interoperability: ASTP/ONC Deregulatory Actions To Unleash Prosperity*. Proposed rule. 90 Fed. Reg. 60970–61034 (Dec. 29, 2025); Doc. No. 2025-23896; RIN 0955-AA09; comments closed Feb. 27, 2026; with the ASTP/ONC HTI-5 Proposed Rule Fact Sheet. federalregister.gov PROPOSED RULE
22. U.S. Food and Drug Administration. *Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions*. Final guidance issued Dec. 4, 2024; reissued Aug. 18, 2025; Docket No. FDA-2022-D-2628. fda.gov GUIDANCE
23. U.S. Food and Drug Administration. *Request for Public Comment: Measuring and Evaluating Artificial Intelligence-enabled Medical Device Performance in the Real-World*. Sept. 30, 2025; Docket No. FDA-2025-N-4203; comments through Dec. 1, 2025. Not guidance. fda.gov GOVERNMENT REPORT
24. The Joint Commission; Coalition for Health AI. *The Responsible Use of AI in Healthcare (RUAIH)*. Released Sept. 17, 2025. Voluntary, non-governmental guidance. jointcommission.org GUIDANCE
25. Miliard M. *Joint Commission intros new voluntary AI responsibility certification*. Healthcare IT News. June 2, 2026. Trade reporting. healthcareitnews.com JOURNALISM
26. Nong P, Adler-Milstein J, Apathy NC, et al. *Current Use And Evaluation Of Artificial Intelligence And Predictive Models In US Hospitals*. Health Affairs. 2025;44(1):90–98. [doi:10.1377/hlthaff.2024.00842](https://doi.org/10.1377/hlthaff.2024.00842) NATIONAL SURVEY
27. Centers for Medicare & Medicaid Services. *ICD-10*. Code update files for FY 2026 (effective Oct. 1, 2025 and Apr. 1, 2026) and FY 2027 (effective Oct. 1, 2026). cms.gov GOVERNMENT REPORT

- 28.** Davis SE, Greevy RA Jr, Lasko TA, et al. *Detection of calibration drift in clinical prediction models to inform model updating*. Journal of Biomedical Informatics. 2020;112:103611.
[doi:10.1016/j.jbi.2020.103611](https://doi.org/10.1016/j.jbi.2020.103611) **METHODS STUDY**
- 29.** Youssef A, Pencina M, Thakur A, et al. *External validation of AI models in health should be replaced with recurring local validation*. Nature Medicine. 2023;29(11):2686–2687.
[doi:10.1038/s41591-023-02540-z](https://doi.org/10.1038/s41591-023-02540-z) **POSITION PAPER**

RELATED PAPERS

Regulating AI in certified EHRs

The certification and device rules behind the monitoring duties described here.

Human oversight of clinical AI

Why an override rate measures alert quality as much as clinician attention.

Evaluating clinical language models

What an accuracy figure needs before it carries information at a single point in time.