ORION FORUM

Before Medical AI Learns on the Job, FDA Should Test Its Teacher

AUGUST 17, 2026

In test environments, medical AI systems have begun to show early signs of the ability to learn during use from feedback they or other models generate. The FDA, through its Predetermined Change Control Plan (PCCP), already has a governance mechanism in place that preauthorizes future updates to medical AI systems under specified conditions. Today, PCCP lacks a requirement that the machine-generated signal teaching these medical AI systems must be tied to real clinical performance. Before the FDA allows these kinds of patient-facing changes to occur in the field autonomously, a medical device's PCCP should identify the signal and technical mechanism doing the teaching, provide evidence that it remains tied to clinical performance, and state the counterfactual, what happens when it does not.

On March 2, 2026, a research team at the Chinese University of Hong Kong published a study introducing an experimental medical AI system called RetExpert. Because medical AI systems can lose accuracy when analyzing patient data different from what they were trained on, RetExpert was designed to adjust itself when retinal images were analyzed from different patient populations, cameras, and clinical settings than the data used to train it. It accomplished this by using its own predictions as temporary labels, called pseudo-labels, and uncertainty scores to decide how much weight to assign to each prediction. The researchers then reset the model after each test batch so any errors would not carry over to the patients that followed. As a result, across 15 unseen clinical datasets, RetExpert outperformed the other AI models tested on disease detection, reliability, and performance with unfamiliar clinical data.

The reset RetExpert took after each batch in testing is a key variable here. RetExpert is a research framework, not an FDA-authorized device or one that learns across patient

interactions. Regardless, it offers an interesting and early example of the kind of adaptive medical AI systems regulators will soon see more of. These new systems raise a distinct regulatory challenge for AI used to diagnose disease, predict risk, triage patients, or guide treatment. It broadens the scope of regulatory oversight beyond the standard non-adaptive medical AI device review that considers whether a model is accurate when the FDA first reviews it to whether the process that teaches the model to change after authorization is robust, accurate and trustworthy.

Congress authorized the creation of PCCPs in December 2022. The FDA then issued its final AI-specific guidance for PCCPs in December 2024 (reissued August 18, 2025). The AI-specific guidance includes recommendations on the information to include in PCCPs for both manual and automatic changes, including "continuous learning" ones, in which software updates after the device is in use. This forward-looking guidance is pro-innovation and nonbinding, but it should not be seen as an industry blank check. Through the PCCP, FDA recommends that companies explain how they will develop and trigger future updates, mitigate risks such as bias and overfitting, test updates on separate data against a clinical reference, set pass/fail criteria, and monitor performance after release.

A device manufacturer can submit a PCCP with a new device it is seeking to bring to market or add one later. Once a PCCP is authorized by FDA, changes covered within it can be made without a new submission for that now newer version. A May 2026 JAMA Health Forum research letter found PCCPs included in 43 of the 794 AI-enabled devices that were authorized from 2023 through 2025, or about five percent (5.4). The final quarter of 2025 saw the share rise to just under ten percent (9.7).

FDA's current guidance already covers a significant amount of the update process, but it does not treat a machine-generated adaptation signal as a distinct entity that requires continuous alignment with clinical outcomes across updates. That signal could be a number of things including a pseudo-label, confidence score, reward model, or learned evaluator (a second AI that grades the first's work). In simple terms, it acts as the medical AI system's teacher. If a confidence score determines which updates proceed then altering that score can change what the device ultimately learns, even if a different test determines whether the version actually

reaches patients.

Recent studies demonstrate this distinction and why it matters. A paper accepted at NeurIPS 2025 aptly titled, TTRL: Test-Time Reinforcement Learning, introduced the novel technique test-time reinforcement learning (TTRL). TTRL is a new method for training a language model that lets it update itself while working on new questions without being shown the correct answers. For each question a model is given, it produces several responses. Its most common answer becomes a temporary answer key and responses matching that answer receive a reward meant as reinforcement. In the study, researchers used the real answers at the end to assess model improvement, but kept those answers hidden during training. Although TTRL improved performance in this study, it demonstrates the key risk: agreement does not always equal accuracy. If a model continually gives the same wrong answer, its feedback system could reward that error and reinforce it. A *Findings of ACL* paper from July 2026 demonstrated this risk, finding that noisy pseudo-labels can create bad learning signals and applying TTRL can make that already bad signal worse.

These findings come from research systems in test settings, not FDA-authorized devices currently on the market, and none of them specifically show that a PCCP has failed. What they show is why a machine-generated teaching signal deserves additional scrutiny before a more challenging or more dangerous case reaches patients.

The governing principle I am proposing to address this is measurement before control. In a previous Orion policy brief on AI and cybersecurity I explored how policymakers were measuring what frontier models could do in laboratory tests while lacking a reliable record of how much AI contributed to real-world cyberattacks. The proposed solution was to make that missing fact more visible by amending an already existing reporting system under the Cyber Incident Reporting for Critical Infrastructure Act (CIRCIA). Medical AI authorization faces a similar problem, just at a different point in its lifecycle. Regulators may see whether one proposed model passes a test, but lack evidence that the process guiding repeated future updates remains tied to clinical truth. FDA should build the measurement layer at the point in the process that creates risk before it expands the control layer.

The FDA does not need to add a new section to the existing PCCP or create a universal technical design for adaptive medical AI devices. When feedback created by a model or another system materially shapes an update in a way that can affect clinical performance, the FDA should require risk-based evidence that the signal is traceable, clinically valid, and subject to appropriate controls. That additional requirement can fit within the existing PCCP sections related to what may change, how those changes will be made and tested, and how risks will be managed.

In the spirit of offering potential solutions, I propose adding some version of the following four questions to the existing PCCP:

- **Question 1**: What creates the learning signal, and what data feeds it?
- **Question 2**: What independent evidence shows that a higher internal score is tied to an intended clinical endpoint across different groups of patients, score ranges, and operating environments?
- **Question 3**: How could the learning signal fail, drift, or be misused, and how will the company discover these problems?
- **Question 4**: What sets of conditions would force the system to pause learning, limit use, trigger human intervention, or revert to a previously approved version?

Additionally, the release safeguard should be proportionate to the kind of change taking place. A new version rolling out to a whole hospital group should undergo a more strict clinical evaluation than a smaller change made at a single hospital or just for one patient. This nuance recognizes that a new clinical evaluation at each step for small changes is not a practical solution. The PCCP should instead explain the guardrails containing those updates, fixed safety monitors, shadow testing when possible, clinician override, routine independent checks, and what safe state it will enter if something goes wrong. A model's self-score should never be the sole justification for the release of an update.

Next, recordkeeping and monitoring can be located inside the quality system manufacturers keep under the Quality Management System Regulation, which covers how a company designs, tests, documents, and corrects its device. As part of this recordkeeping, companies

should maintain privacy-compliant records detailed enough for regulators to see the decisions behind each model change and release. Finally, any change that occurs out of scope of the authorized PCCP should go through normal FDA change assessment or, if needed, a new marketing submission.

The proposed solution also offers a response to questions FDA has been asking. In September 2025, the FDA sought public comment on measuring and evaluating real-world AI-enabled medical device performance, including model drift, device logs, monitoring triggers, clinical outcomes, and user feedback. Section 515C of federal law permits the FDA to require performance requirements for PCCP changes. FDA can also release recommended evidence in nonbinding guidance and ask for device-specific assistance during the review process before a device is marketed. This proposal does not require a new law, authority, or office to implement and it does not require a ban or pause on adaptive learning.

Finally, the new testing requirement should apply only when learned or self-generated feedback has a meaningful effect on a change that can affect clinical performance or the device's release path. If removing the signal does not materially affect those, then it does not meet the threshold to trigger this new added review. In those cases, the standard FDA change-control process would still apply.

PCCPs answer a key question: How can the FDA approve future changes to medical devices without the need to review each iterative version from scratch? Adaptive learning brings the need to add a follow up: What taught the medical AI system that a specific change was better? Today, the FDA evaluates the student and looks at the lesson plan. It should extend that testing to the signal and the teaching system. The goal is not to stop medical AI from learning, but to ensure that the system's teacher is not the only part of the loop that goes untested.

OPI.