UNLEARN

June 29, 2026

Division of Dockets Management (HFA-305)
Food and Drug Administration

Docket No. FDA-2026-N-4390: AI-Enabled Optimization of Early-Phase Clinical Trials Pilot Program; Request for Information (91 FR 23100)

Dear Deputy Chief Medical Officer Mundkur and FDA Pilot Program Team:

Unlearn.AI appreciates the opportunity to respond to FDA's Request for Information on the AI-Enabled Optimization of Early-Phase Clinical Trials Pilot Program. We support the Agency's view that early-phase trials are a critical bottleneck in drug development and that AI-enabled technologies can materially improve the efficiency, speed, and quality of decision-making in this setting. We offer the comments below to help the Agency design a pilot that is rigorous, trustworthy, and beneficial.

Our comments are organized using FDA's question numbering. We would welcome the opportunity to discuss these points further. Please direct questions to Jonathan Walsh, Chief Scientific Officer (jon@unlearn.ai).

Respectfully,

Jonathan Walsh,
Chief Scientific Officer
Unlearn.AI

About the respondent

Unlearn.AI is a technology company that has used AI to help sponsors design and analyze early and late phase clinical trials since its founding in 2017. The company uses historical patient data to create AI-based digital twins, comprehensive, probabilistic predictions of participant outcomes on a defined standard of care. Through these technologies, Unlearn.AI helps sponsors run simulations to plan trials and to increase the efficiency of analyses. This includes the development and advancement of prognostic covariate adjustment (PROCOVA) [EMA opinion, FDA comments], methodology to utilize prognostic models to add statistical power or prospectively reduce sample size in randomized controlled trials.

Unlearn is not a sponsor, but works with sponsors on the above activities, independently builds AI models, and works to advance regulatory-compatible use cases of AI in drug development. As a technology company, we are driving and benefit from the increased adoption of AI in clinical development, in particular the use of simulation-based methods and digital twin models to inform new decision-making paradigms such as those in this Request for Information and associated Pilot Program.

Executive Summary

Unlearn.AI appreciates the opportunity to respond to FDA's Request for Information on the AI-Enabled Optimization of Early-Phase Clinical Trials Pilot Program. We support the Agency's view that early-phase trials are a critical bottleneck in drug development and that AI-enabled technologies can materially improve the efficiency, speed, and quality of decision-making in this setting.

This response focuses on answering Question A.1.c. about AI use cases, and also provides answers to several other questions collectively. We propose two complementary use cases, both grounded in machine-learning digital twin models:

1. **Digital twin efficacy benchmarks**. Digital twin models can generate participant-level predictions of how each enrolled patient would have progressed under a range of control conditions, providing richer, population-matched benchmarks. They are particularly valuable in early-phase trials where concurrent controls are absent or small. The same predictions can be used to flag potential operational or data anomalies, including site-specific deviations, earlier in the trial, so that remediation strategies can be initiated sooner.

2. **Real-time trial simulation.** The same modeling infrastructure can serve as a real-time simulation engine: as Phase 1 and early Phase 2 data accrue, the model is progressively recalibrated and used to simulate candidate designs for the next phase before the current phase has concluded, increasing confidence of recommendations, and potentially shrinking the gap between phases.

These applications can lead to better decision-making that can trigger the next phase of trials earlier, shorten development timelines, provide clearer evidence to regulators and sponsors to contextualize trial results, and ultimately get effective treatments to patients sooner. In our collective response, we provide recommendations on the conduct and evaluation of the Pilot Program from the perspective of an organization increasing adoption of new, AI-driven technologies in clinical development.

AI use cases in early stage trials

This section is a response to question A.1.c from the RFI.

In general, there are many uses of AI that can benefit early stage clinical trials, such as enabling single arm trials or other efficient analyses that lead to smaller enrollment. However, many of these are actions that sponsors may take and engage with the agency through regular processes. We focus on use cases in the spirit of real-time evaluation of clinical trial data that benefit from more active agency involvement and allow robust decision-making by sponsors in cooperation with the agency, while still enabling rapid early phase timelines.

We propose two use cases based on machine-learning-generated digital twins which aim at improving trial efficiency and decision quality: Digital Twin Efficacy Benchmarks and Real-Time Trial Simulation. The goal of the first is to set various benchmarks that allow for more fine-grained evaluation of ongoing trial data, allowing for earlier decisions on clear signs of efficacy, greater contextualization of trial results, and detection of data quality issues. The goal of the second is to provide a concurrent link between the conduct of a study and planning for the next phase, to shorten the time between phases and allow for joint, program-level decision-making.

Digital Twin Efficacy Benchmarks

Limited benchmarks in early-phase trials

Many early-phase trials (e.g., in rare diseases and oncology) are characterized by limited or absent concurrent controls. Most Phase 1 studies are single-arm; many Phase 2 studies include only a limited concurrent control. Sponsors and reviewers are therefore typically limited in the interpretation of early signals against fixed historical benchmarks or against noisy estimates of the control-arm trajectory. Both comparators are challenging: fixed benchmarks are not tailored to the specific population enrolled, and small concurrent controls produce noisy estimates of the control-arm trajectory. This is the natural target for AI-enabled, model-based comparators.

Digital Twins

By digital twin we mean a participant-level computational model that, conditional on a participant's baseline characteristics (demographics, disease history, labs, imaging-derived features, genomic markers, prior treatment, etc.), produces a prediction of that participant's trajectory under one or more counterfactual control conditions (e.g., standard of care A, standard of care B, best supportive care). Digital twins are generated with machine-learning models trained on historical control and observational data.

The role of benchmarks

In this use case, we are not proposing that digital twins replace concurrent randomized controls or confirmatory statistical analyses where those are appropriate. Their role instead is to provide higher-quality, individualized supportive evidence, delivered in near real time, that complements pre-specified analyses and informs internal and regulatory decision points earlier than otherwise possible. In this setting, the use of AI contributes to decision quality and timeliness and not to formal statistical inference about treatment effect, which aligned with FDA's goal of enabling "more informed early go/no-go decisions" while preserving rigorous scientific and regulatory standards.

The unique advantage of model-based comparators

A key advantage to use machine-learning-based digital twins is that modern architectures can synthesize heterogeneous evidence at a level traditional comparators cannot:

- Multi-source: individual-patient-level data from prior randomized control arms (sponsor and consortium trials).
- Heterogeneous: real-world data from EHRs, registries, and claims.
- Multimodal: imaging, omics, structured labs, and free-text clinical notes.
- Multi-resolution: aggregate summary statistics from published trial reports alongside individual-level records.

This synthesis makes the resulting benchmarks more robust and enables generation of multiple counterfactuals.

Early anomaly detection

A useful by-product of producing per-participant counterfactual predictions is the ability to compare each participant's observed trajectory against the model's expectation in near real time. Systematic, unexpected deviations – particularly when concentrated at a specific site, investigator, batch, or vendor – can serve as an early warning of operational or data-quality anomalies, including:

- protocol deviations or dosing errors at a specific site;
- assay or laboratory drift affecting biomarker readouts;
- site-level enrollment drift (e.g., systematically different baseline severity than the rest of the trial);
- data entry or coding inconsistencies.

We propose this as an optional component of the pilot, layered on top of the primary counterfactual-benchmark use case. The same per-participant predictions used for supportive efficacy evidence are aggregated by site (or other operational stratum) and monitored for systematic departures from the joint distribution implied by the digital twins. When anomalies are found, the flags would be advisory, used to trigger conventional quality follow-up (monitoring visits, source data verification, central review), and would not be used to adjudicate participant data or treatment effects directly. Anomalies that would normally surface only at database lock or interim analysis can be detected and remediated weeks or months earlier, reducing the risk of affecting downstream decisions.

Expected impact

Richer, tailored benchmarks support earlier and better decisions:

- Patient impact: effective drugs are available sooner.
- Joint impact: ability to interpret trial data and ensure data quality in the context of model-based, comprehensive simulations via digital twins.
- Regulatory-relevant impact: improved quality and timeliness of go/no-go decisions, reducing late-stage failures attributable to underpowered or misinterpreted early signals.
- Sponsor-relevant impact: shortened timelines between phases, since sponsors can begin preparing the next phase earlier when AI-derived supportive evidence converges on a confident read sooner than the pre-specified primary analysis.

Real-time trial simulation

Concept

The second use case extends the first from digital twins of each participant to digital twins of the trial:

- Begin with a pre-trained digital twin model.
- As the Phase 1 (or early Phase 2) trial enrolls, ingest accruing control- and treatment-arm data in near real time and use them to recalibrate and fine-tune the model. Treatment-arm data inform a treatment-effect component layered on top of the control-arm twin.
- The resulting time-evolving model serves as a live in-silico simulation engine for the design of the subsequent phase.

Trial data continuously feed the simulator, and the simulator continuously informs the design of the next phase.

Design optimization

The most immediate application is to use the live simulator to optimize Phase 2 design while Phase 1 or Phase 2a is still ongoing. Design questions that can be addressed in silico include:

- Population. Which subgroups appear most likely to benefit, and how should inclusion/exclusion criteria be tightened or broadened?
- Endpoints. Among candidate endpoints (e.g., ORR vs. PFS vs. biomarker composites), which best trade off sensitivity, time-to-readout, and regulatory interpretability?
- Duration. What follow-up duration is sufficient to detect a clinically meaningful effect, given simulated control trajectories?
- Sample size and randomization ratio. What design produces acceptable operating characteristics under realistic effect-size and heterogeneity assumptions, for different endpoints? Is the proposed design appropriate?

All of these can be addressed using a simulator informed by both historical data and the current trial – making the Phase 2 design materially better-calibrated to the actual population and treatment behavior observed in earlier phases.

Impact

A live digital twin of the trial directly serves two key goals of the pilot:

- Efficiency and speed: it compresses the time between phases by allowing next-phase planning to begin while the current phase is still maturing.
- Decision quality: it grounds design choices in a model calibrated to both historical data and the specific trial population.
- Consensus: common information on study design helps the agency and sponsor understand and align on study feasibility.

General considerations for the use of AI for real-time decision-making in clinical development

In this section we provide broader comments that address several of the questions in the RFI. In several cases we label where commentary addresses particular questions.

As a broad discipline, clinical trials are one of the largest research and development investments made by society, and one that has perhaps the largest impact on individuals. In the age of AI, patients stand to benefit from rapidly expanding capabilities to better manage the complex infrastructure and decision-making throughout clinical trials, lessening the burden on patients and bringing effective treatments to market faster. Innovation, driven by FDA as perhaps the single most important participant in the drug development process, is vital to shape the paths for decision-making in drug development and to achieve the goals of benefits for patients. FDA and our community broadly can take inspiration from other large scale experiments such as the Large Hadron Collider, which by necessity due to data complexity have long utilized simulations and machine learning tools to make real-time decisions and generally in the planning and interpretation of experiments (see, e.g., this review or this review).

The Pilot Program associated with this initiative will take an important step in creating new pathways for decision-making. As such, thoughtful selection of the pilots allows the agency to clearly evaluate benefits and build trust among stakeholders. *[Question A.1.a]* In general, we believe that trials where safety is less of a concern (such as with established modalities with well understood safety profiles or with targeted therapies) provide better trials for the application of AI for decision making. For example, if a pilot study uses a clear signal of efficacy as a trigger for acceleration to a later phase, the benefit is limited if the study must continue running after the efficacy to accrue adequate safety data to initiate the next phase. The most likely trials to benefit would be (1) Early phase studies, where no concurrent control or a limited control group exists, (2) Phase 1b/2 oncology cohort expansions and Phase 2 proof-of-concept studies in indications with heterogeneous populations and evolving standards of care, or (3) Rare-disease early-phase trials, where small samples and limited natural-history data make higher-quality comparator evidence especially valuable. *[Question A.1.b]* The selection of therapeutic areas should be broad, within the considerations above. That being said, clear signals of success are important to build credibility for the effort, which may encourage selectivity to favorable study designs.

In the spirit of promoting successful piloting and maturation of a technology system (AI and associated tools) that we strongly believe in, and speaking from experience in helping drive

adoption of an AI-based method for more efficient clinical trials, prognostic covariate adjustment (PROCOVA), we provide commentary on the steps that the agency can take on the structure of the Pilot Program.

We view the broad, deep adoption of AI tools within clinical development as inevitable, driven by the demonstrated value from existing use cases and the growing use across industries. One may regard the increase of real-time decision-making during clinical development, in participation with regulators, as similarly inevitable. We can expect future infrastructure to exist where data management of studies is straightforward, interoperability concerns are low, and there are capabilities to contextualize ongoing study data through the use of historical data, AI tools, and simulations. In this case, real-time decision-making is quite feasible, and there will be pathways for effective drugs to have accelerated paths through clinical development and to empower regulators to act efficiently in the review process. This necessarily requires the participation of regulators, such as in this Pilot Program.

Therefore, a natural position to take is that the endpoint is inevitable, and we must chart a course that balances responsibility and expediency. Rash change that jeopardizes trust will slow or stop progress. But active change is needed to encourage adoption, since a slow path that waits for the clinical trial ecosystem to naturally adopt new technologies can take decades.

[Question A.4.a] The structure of the Pilot Program should allow the agency to identify approaches to real-time decision-making while also supporting alignment with sponsors. To that end, FDA should provide support through upfront engagement to define criteria for interpreting data against benchmarks or for the refinement of future study designs based on real-time information. This will create clarity with sponsors on success. Similarly, FDA should provide both expectations for what data will trigger actions by the agency and mechanisms for meetings with sponsors to evaluate ongoing study data. *[Question A.4.c]* For technical guidance, FDA should expect varying degrees of technical maturity across participants, and should not disqualify sponsors based on their AI capabilities or experience. The agency can support such sponsors by providing deeper recommendations for such pilot participants and use it as an opportunity to calibrate the breadth of monitoring and consequently the set of actions taken.

A graduated set of approaches is necessary for new technology adoption, especially when experience varies extensively across the field. *[Questions A.4.b, B.4.a, B.4.b, B.5.a, B.5.b]* This is also reflected in the infrastructure used for the Pilot Program. While it is tempting to utilize a single, consistent architecture to simplify execution of the Pilot Program, it is also an opportunity for the agency to see differing implementations and understand what tools and approaches are effective. We recommend providing a minimal set of expectations for program participants and to consider updating those expectations to be more rigorous upon expansion beyond the Pilot Program. Beyond the expected security and compliance requirements for shared infrastructure for the transmission of data, we recommend utilizing the framework defined in the draft guidance on AI in drug development, *Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products,* as a way to align with sponsors on the use of models and the risks expected in Pilot Program. This will also naturally suggest what shared tools FDA should have access to. The model credibility assessment framework naturally lends itself to a clear identification of the use cases of AI models, their risks, the plan for deployment, and metrics to assess credibility. In the use cases we propose, metrics evaluating calibration and discrimination of model predictions can be evaluated on historical data (both for control benchmarks and simulated designs). If relevant data is collected during the trial (e.g., concurrent controls) it can also be used to evaluate model performance.

The outcomes and reporting of the Pilot Program will be important to establish trust with stakeholders. *[Questions A.6.a, A.6.b, B.1.a, B.1.b, B.7.a, B.7.c]* As such, we recommend a clear set of metrics to evaluate studies during the Pilot Program, disseminated through whitepapers and public meetings. Participating sponsors and the agency should agree upon expected information in-bounds for public dissemination at the outset, both in any outcome of the program and subject to discussion after study results are known. We recommend the FDA consider writing a paper on the Pilot Program after completion that evaluates successes and challenges – both for operations and scientific output – in a blinded way, avoiding attribution of characteristics to individual programs yet allowing the community to understand the landscape of activities. Reports can also assess the prospective value of real-time decision-making, the steps required for operational scalability, and the need, where appropriate, for technical maturation among stakeholders to increase the benefit of real-time decision-making. Public meetings such as workshops can be used to evaluate stakeholder trust in approaches and help shape follow-up activities. As an example for metrics, in the use cases proposed, the expected benefit is to reduce the time between Phase 1 and Phase 2. Useful metrics in this case are (1) the calendar time from last-patient-last-visit in Phase 1 to first-patient-in in Phase 2, (2) the calendar time from interim/final Phase 1 readout to Phase 2 protocol finalization. In pilots using live digital twin design optimization, another useful metric is an estimate of time saved by parallelizing next-phase design against the maturing trial. We caution that timelines are confounded by many non-AI factors and recommend pairing them with case-by-case qualitative attribution. *[Questions B.2.b, B.2.c]* Metrics should also compare to the expected development timelines through traditional pathways for similar programs and designs, and anecdotally identify differences in study conduct and the changes in decision-making, including those due to AI. We emphasize that metrics can be quantitative, but the overall experience should be placed in context qualitatively, and contrast to traditional approaches can be quite effective in evaluating value.

Ultimately, as with any new technology, one should expect results and stakeholder satisfaction to be heterogeneous and improvements necessary both in operations and organization. Careful design of the Pilot Program can subvert many challenges, but resolute leadership is needed to lead transformation and bring forward new technologies.

Conclusions

We thank the FDA for proposing this pilot and for the opportunity to comment. A focused pilot on digital-twin-based counterfactual benchmarks and live trial simulation can produce both concrete near-term gains in early-phase decision quality and a durable methodological foundation for the broader use of AI in clinical development.

Unlearn.AI would be glad to engage further with FDA on any aspect of this response, including additional technical detail, retrospective case studies, or methodological references.

Respectfully,

Jonathan Walsh
Chief Scientific Officer
Unlearn.AI
jon@unlearn.ai