

Response to Request for Information:

AI-Enabled Optimization of Early-Phase Clinical Trials Pilot Program

Docket No. FDA-2026-N-4390 | Submitted May 28th, 2026

Submitted by: Veronica Vu, VP of Data at Lindus

Contributors: Michael Young, Co-CEO; Gemma Harrison, Chief Product and Technology Officer; Shini Pattni, Head of Legal; Collin Anderson, Clinical Project Lead

Framing Statement

We submit this response as an end-to-end clinical development organization. We develop AI tools and deploy them in the trials we operate as a CRO, and also are developing our own portfolio of therapeutic assets. We are one of a small number of organizations that would hold both sponsor and CRO accountability for AI deployment in live trials simultaneously. We are not advising on what sponsors should do with AI tools we have built for them. We would be subject to the same regulatory obligations as any sponsor, and we are operationally responsible for the infrastructure those tools run on. Under ICH E6(R3), we function as a service provider; under the NIST AI Risk Management Framework, we function as developer, deployer, and operator within the same program. That gives us a different view of where AI succeeds and fails in practice than either a pure vendor or a pure CRO would have.

We support FDA's proposed pilot. Five observations, grounded in that combined experience, run through our responses below.

- Data pipeline readiness is a prerequisite for AI performance, not a variable to be accommodated. A model's technical capability cannot compensate for data that arrives late, incomplete, or inconsistently formatted. Evaluation frameworks that conflate pipeline performance with model performance will produce findings that cannot be interpreted cleanly.
- Organizations that carry both sponsor and CRO accountability for AI deployment occupy all three roles the NIST AI RMF defines: developer, deployer, and operator. The pilot's governance structure does not currently account for this. That gap needs to be addressed explicitly, because these organizations are among the few that can provide evidence about how AI performs across the full deployment lifecycle under regulatory accountability.
- To determine whether AI-generated insights have value in early-phase trials, AI tools need to run in parallel with full trial execution, generating and recording outputs at defined intervals without those outputs driving decisions in real-time. Those recorded outputs are then compared against the trial's final results. Repeated across multiple trials and datasets, that comparison is what produces evidence about whether early AI insights are predictive of outcomes that matter. The pilot's value depends on that mechanism being built in from the start, not added retrospectively.
- AI maturity is multidimensional. A sponsor may be advanced in modeling capability but nascent in data governance, or organizationally ready but deploying a tool with limited validation evidence. Treating maturity as a single dial will produce a pilot that fails to surface the specific gaps that most commonly drive deployment problems.

- Explainability requirements differ by stakeholder in kind, not just degree. What an investigator needs to act on a safety alert, what a data safety monitoring board needs to evaluate model performance, and what a regulator needs to accept AI-assisted analysis are structurally different information needs. A single explainability standard applied uniformly will not serve any of them well.

Section A: Pilot Program Design and Implementation

A.1 Scope and Focus

A.1.a: Which trial types or trial issues might benefit most from AI?

Oncology dose-escalation studies would benefit particularly from AI-assisted integration of pharmacokinetic, pharmacodynamic, and safety data ahead of escalation committee meetings, where current practice involves manual synthesis that is time-consuming and can introduce inconsistency across analysts. This is especially relevant as sponsors develop increasingly complex modalities, including siRNAs, bispecific antibodies, and gene therapies, where quickly identifying the optimal biological dose is commercially critical for small and mid-size sponsors operating with limited capital and narrow development windows. Rare disease trials present a different case: participant recruitment timelines are often a primary driver of trial duration, and AI-assisted eligibility screening against structured electronic health record data has the potential to compress those timelines meaningfully. The relevance extends to adaptive designs, which are becoming more common in early-phase trials and where AI's ability to synthesize accumulating data in near real-time is well-matched to the decision cadence those designs require.

We recommend against prioritizing biomarker-based endpoint validation in the initial pilot. The data infrastructure requirements are more demanding, and the regulatory precedent for AI-derived endpoints is still developing. Including it risks overloading the pilot with questions it cannot yet answer cleanly.

A.1.b: Should the pilot target specific therapeutic areas or remain broadly applicable?

The pilot should remain broadly applicable within reason, so that FDA and industry can evaluate which therapeutic areas show the most benefit from AI-assisted approaches. That said, the earliest trials selected should anchor in areas where FDA has the most data history and regulatory precedent, with oncology and rare disease as strong candidates. A structured breadth approach, where initial slots prioritize areas with established precedent while later slots open to sponsor proposals, balances the need for meaningful early comparison against the value of generating evidence across a wider range of settings.

A.1.c: Should priority be given to specific AI use cases?

Yes. In approximate order of feasibility and near-term impact: safety signal detection and adverse event triage, where AI can augment human monitoring without replacing clinical judgment; dose selection support in escalation studies; participant eligibility screening against structured data sources; and protocol deviation detection. These use cases share a

characteristic that makes them well-suited to early piloting: the value of getting them right is clear, and the cost of getting them wrong is detectable.

A.2 Participant Selection

A.2.a: What criteria should FDA use to select sponsors, trials, or technologies?

Selection criteria should assess three dimensions independently rather than as a composite score: data governance maturity, covering whether the sponsor has documented data quality standards, clear data lineage, and defined processes for handling missing or inconsistent data; trial operational readiness, covering whether the protocol is finalized, investigator locations identified, and data collection infrastructure in place; and AI tool validation status, covering whether the tool has been validated on data representative of the intended trial population with failure modes documented.

Conflating these into a single readiness score will advantage sponsors whose documentation is strong over those whose infrastructure is strong. FDA should consider a structured intake process that assesses each dimension separately. Critically, these criteria should determine the level of support and monitoring a sponsor receives in the pilot, not whether they are eligible to participate. A sponsor who is nascent on one dimension should qualify for the foundational tier with structured support, not be excluded.

We recommend FDA add a fourth dimension to the intake assessment: data infrastructure integration. Specifically, whether the AI tools being deployed operate within an integrated data infrastructure that governs data collection, management, and model inputs as a unified system, or whether they are bolted onto external EDC systems after the fact. AI tools that depend on external EDC data feeds inherit whatever latency, inconsistency, and format variability those systems produce. An integrated architecture, where data collection and AI operation share the same infrastructure, eliminates an entire class of failure modes before the model runs. FDA should weigh this as a positive selection criterion.

A.2.b: How can the pilot ensure representation across organization size, capability, and therapeutic area?

Set explicit targets for each dimension and treat them as constraints rather than aspirations. A pilot dominated by large sponsors with mature data science teams will not generate findings generalizable to mid-size or emerging sponsors. Reserve a defined number of slots for sponsors below a threshold of trial volume or organizational scale, and pair them with structured technical support rather than requiring them to meet the same entry bar as larger organizations.

A.3 Collaboration Models

A.3.a: What partnerships are most effective?

The RFI names sponsors, technology vendors, academic partners, patient advocacy groups, and FDA as stakeholders in the pilot. CROs are not included. In many early-phase trials, the CRO owns data collection, manages relationships with investigator locations, and operates the data acquisition tools. Any AI tool running on trial data is running on infrastructure the CRO manages. FDA should name CROs explicitly in the pilot's governance structure.

For technology-enabled CROs like ours, there is an additional consideration. We develop the AI tools and deploy them in the trials we manage, which means we hold developer, deployer, and operator responsibilities under the RMF simultaneously. The pilot's collaboration model would benefit from accounting for organizations that span all three roles, as each carries distinct accountability obligations that are all relevant to how the pilot is governed.

A.3.b: How can FDA facilitate pre-competitive collaboration and knowledge sharing?

A structured working group convened by FDA at defined intervals, where participants can share de-identified process learnings without disclosing proprietary data, is the most practical mechanism. FDA should publish the framework for what can and cannot be shared before the pilot begins. Retrospective publication of aggregated findings is valuable but insufficient on its own; real-time sharing of implementation challenges is where the most actionable learning happens.

A.3.c: What role should patient groups and investigators play in AI governance?

Patient advocacy organizations should have input into how AI tools used for participant screening or recruitment outreach are designed, not solely in reviewing outputs after the fact. Investigators should participate in AI governance at the investigator location level, including training on how AI-generated alerts are meant to be interpreted and acted upon. Investigators who participate in AI governance at the location level (including training on how alerts are meant to be interpreted and acted upon) are better positioned to use AI outputs effectively and flag when those outputs are not calibrated to clinical reality.

A.4 Operational Structure

A.4.a: What support should FDA provide?

Three categories: regulatory clarity on what AI-assisted data or decisions will and will not be accepted in submissions arising from pilot trials, issued before the pilot begins; a defined point of contact within FDA for AI-specific questions as they arise during trials, rather than routing through standard review channels; and technical guidance on minimum documentation standards for AI tools used in the pilot, translated into practical requirements for clinical trial contexts. On the first point: upfront clarity on what FDA will accept from AI-assisted submissions meaningfully lowers the participation risk for sponsors, particularly smaller organizations that need to know the regulatory basis for their approach before committing to a trial design built around it.

A.4.b: What infrastructure is needed?

The infrastructure gap with the greatest bearing on AI performance is not compute or model access but the relationship between data collection and AI operation. FDA should require, as a condition of pilot entry, that AI tools operate within an integrated data infrastructure rather than as add-ons to external EDC systems. The distinction is material: a bolt-on architecture means the AI tool receives data after it has passed through a separate collection system, inheriting that system's latency, format inconsistencies, and data quality gaps. An integrated architecture, where data collection and AI operation share the same underlying infrastructure, means the model operates on data governed by the same standards that produced it. This is not a

preference for a particular vendor or platform; it is a structural requirement for the data pipeline reliability the pilot is trying to evaluate. It also reflects how we have built our own infrastructure; data collection and AI operation are not separate systems that communicate with each other, but a single environment governed by the same standards throughout.

FDA should treat this as a GOVERN activity under the RMF: a precondition for meaningful participation, not a variable to be accommodated after enrollment. Sponsors using bolt-on AI tools should not be excluded from the pilot, but their results should be analyzed as a separate cohort, because the infrastructure question is itself one of the most important things the pilot can generate evidence about.

A.4.c: How can the pilot accommodate varying levels of AI maturity?

The RFI frames AI maturity as a spectrum from low to high, but it is better understood as several independent dimensions. A sponsor may have sophisticated modeling capability but weak data governance. Another may have strong infrastructure but limited experience translating model outputs into clinical decisions. These sponsors need different kinds of support, and a single maturity score will not surface that distinction.

The dimension we would most encourage FDA to assess is documented production deployment in live trials, weighted above claimed modeling capability. A tool that has been validated in a development environment and presented with strong technical documentation is not the same as a tool that has operated against real trial data at real investigator locations and has a track record of performance under those conditions. Organizations without production deployment experience will market modeling capability because it is what they have. We speak to this from both sides as the organization that develops and deploys these tools. The pilot should be designed to distinguish between the two, because the failure modes that matter most in clinical trials are deployment failures, not development failures.

We recommend FDA assess maturity across four dimensions at intake: data readiness, model validation, organizational change management, and AI lifecycle management. The fourth dimension, covering whether the sponsor has processes for ongoing monitoring, drift detection, version control, and model updates in a live trial environment, is increasingly recognized in regulatory guidance as what distinguishes basic adopters from organizations genuinely prepared for production deployment. The results of that assessment should determine the level of support and monitoring each participant receives. A foundational tier with structured support and closer monitoring would serve sponsors deploying AI in a live trial setting for the first time; an advanced tier would apply to sponsors with documented prior deployment experience. This approach allows the pilot to generate learnings across the maturity spectrum without setting a single entry bar that would exclude less experienced participants.

A.5 Timeline and Milestones

A.5.a: What is an appropriate duration for the pilot?

A minimum of 36 months. First-in-human studies frequently take 18 to 24 months to reach the dose decisions where AI's contribution to decision quality is most relevant. A shorter pilot will capture onboarding challenges but not whether AI tools improve outcomes at the moments that matter most.

A.5.b: What interim milestones or checkpoints should be included?

Quarterly operational check-ins focused on data quality and pipeline integrity, distinct from formal regulatory reviews. The most common point of failure in AI-assisted trials is not model performance but data pipeline integrity: missing data, inconsistent formats, and delays in DAT entry that degrade AI inputs before the model runs. Quarterly check-ins give FDA visibility into these issues before they become trial-level problems. Formal milestones should include an initial data quality assessment at trial outset, a mid-trial review of AI tool performance against pre-specified metrics, and a final evaluation combining quantitative outcomes and qualitative stakeholder assessment.

A.5.c: How should FDA balance rapid insights with rigorous evaluation?

Define a small set of pre-specified metrics before the pilot begins and evaluate against those rigorously, while allowing observational data collection on a broader set of outcomes. Three to five well-defined metrics mapping directly to trial efficiency, decision quality, and safety monitoring will produce more credible and actionable findings than trying to evaluate everything at once. This focused approach also sets the foundations for broader evaluation in subsequent phases, once the pilot has established which measurement approaches are operationally feasible and which AI contributions are most consistently detectable.

A.6 Knowledge Sharing***A.6.a: How should lessons learned be captured and disseminated?***

FDA should assign a dedicated knowledge management function to the pilot, not treat dissemination as a post-hoc activity. This means structured documentation of implementation decisions and adaptations in real time, using a standardized template all participants contribute to. Aggregated, de-identified versions should be published on a rolling basis rather than held until the pilot concludes.

A.6.b: What mechanisms can promote transparency while protecting proprietary information?

A tiered disclosure model. Process learnings, covering how data pipelines were structured, what monitoring cadence was used, and what integration challenges arose, can generally be shared without competitive risk and should be shared openly. Performance data on specific AI tools should be shared in aggregate or anonymized form. FDA should also consider establishing a regulatory sandbox for the pilot, providing a defined space where participants can test approaches and share learnings with reduced risk of those learnings being used against them in future submissions. For data submissions, FDA should allow participants to submit two versions, one containing full proprietary detail and one with sensitive information removed, consistent with FDA's approach to other submission types, so that the submitting organization controls what is made public rather than having FDA make that determination unilaterally.

Section B: Evaluation Metrics and Success Criteria

B.1 Trial Efficiency and Speed

B.1.a: How should improvements in trial efficiency be measured?

By phase of trial activity rather than at the aggregate level. We recommend measuring separately: time from protocol finalization to investigator location activation, time from activation to first participant enrolled, time from data lock to go/no-go decision, and time from Phase 1 completion to Phase 2 IND filing. Each sub-interval is affected by different factors, and AI tools are likely to have differential impact across them. Results should also be stratified by endpoint type and therapeutic area to ensure meaningful comparison across the pilot's participant cohort. Aggregate cycle time metrics are useful for headline comparisons but mask where time is actually being saved or lost.

B.1.b: What metrics should assess reductions in time from Phase 1 to Phase 2 initiation?

Decision-ready time warrants particular attention as a metric: the point at which sufficient data has been reviewed and interpreted to support a go/no-go recommendation, regardless of when the formal decision is documented. AI tools that accelerate data integration may compress decision-ready time significantly even when calendar time is constrained by regulatory review schedules. FDA should consider how to capture this distinction in its measurement framework.

B.1.c: How can improvements in participant screening, recruitment efficiency, and retention be quantified?

Screen failure rate is a useful starting point for recruitment measurement because it is directly addressable by AI and measurable at baseline without any AI tools in place. Eligibility criteria in early-phase trials are frequently defined by parameters that can be evaluated prospectively against structured patient data: laboratory values, performance status, prior treatment history, and diagnostic findings. AI-assisted screening can apply these criteria systematically before referral, compressing the gap between patient identification and eligibility determination and reducing the rate at which participants consent and then fail to enroll. Metrics should include screen failure rate, time from referral to eligibility determination, and dropout rate in the first 30 days of participation.

B.2 Decision Quality

B.2.a: How can the quality and timeliness of go/no-go decisions be evaluated?

FDA regulatory decisions about whether a study may proceed and sponsor-internal portfolio decisions about whether to continue investment are different problems with different inputs, stakeholders, and appropriate success metrics, and the RFI conflates them. For FDA regulatory decisions, quality metrics should focus on completeness and consistency of the data package and elapsed time between data availability and decision documentation. For sponsor-internal decisions, concordance between AI-supported recommendations and expert panel judgment is a practical starting point. Evaluating AI-supported decisions against pre-selected historical comparator studies matched on endpoint type and therapeutic area would provide a contextually anchored benchmark, though this approach requires tight matching criteria and

careful accounting for changes in standard of care, site experience, or data collection practices over time. Tracking whether Phase 1 go/no-go decisions made with AI support are validated or reversed by Phase 2 outcomes offers the strongest long-term signal, but requires a follow-up window longer than the pilot itself.

B.2.b: What methods can assess concordance between AI-supported and traditional decision-making?

Parallel review designs, where AI-generated recommendations and traditional committee recommendations are generated independently and then compared, offer the greatest methodological rigor. Where parallel review is not feasible, structured retrospective review of cases where AI and human judgment diverged provides valuable learning, particularly if the divergence can be traced to specific data inputs or model assumptions.

B.2.c: How should reductions in late-stage trial failures be measured?

This is a valid long-term outcome measure but not a practical short-term pilot metric. Phase 3 failure rates are influenced by too many factors outside early-phase decision quality to attribute causally to AI tools within a pilot timeframe. Measurement is further complicated by reporting bias: failed trials are less likely to be published, which means any analysis of late-stage failure rates will systematically undercount failures and overstate the apparent contribution of early-phase interventions. We recommend FDA treat this as a long-term tracking commitment rather than a pilot success criterion, build the data infrastructure now to enable that tracking over a five to ten year horizon, and account explicitly for reporting bias in the analytical framework.

B.3 Participant Safety and Data Integrity

B.3.a: What metrics should evaluate safety signal detection and response time?

There is a distinction between model-layer performance and data-layer readiness that the pilot's evaluation framework should reflect. An AI safety monitoring tool that can detect a signal within hours of data entry is limited by the timeliness of the underlying data. In our experience operating both sides of the data pipeline and consistent with broader industry observation, organizations widely lack effective mechanisms for tracking and managing the timeliness of investigator location data entry into DATs. The commonly cited benchmark is 24 to 48 hours after a participant visit, but actual performance varies substantially across investigator locations and studies, and that gap is rarely surfaced in trial-level reporting.

Metrics should therefore be structured in two layers, and we recommend FDA treat this two-layer structure as a core pre-specified requirement rather than an optional reporting format. Data pipeline metrics should include: time from clinical event to DAT entry, completeness rate at scheduled data entry intervals, and data reconciliation discrepancy frequency across investigator locations. Model performance metrics should include: time from data availability to alert generation, alert positive predictive value, and false negative rate on retrospectively identified safety events.

Reporting these separately is not just good measurement practice; it is a structural test of the infrastructure architecture. AI tools operating within integrated data infrastructure should consistently outperform bolt-on tools on data pipeline metrics, regardless of relative model

sophistication. If the pilot does not separate these layers, that signal will be invisible in the results. We recommend FDA require all participants to report both layers from the outset, and analyze integrated versus bolt-on architectures as a primary stratification variable.

B.3.b: How should the impact of AI on adverse event rates or protocol deviations be assessed?

Adverse event rates are a participant safety outcome, not an AI performance metric. AI tools do not reduce adverse events; they may improve the speed and accuracy of detection and the appropriateness of clinical response. The relevant metric is response appropriateness: when a safety signal is flagged, was the clinical response timely and consistent with protocol? Protocol deviation rates are more directly attributable to AI tools when those tools are used for compliance monitoring, and deviation rate trends before and after AI implementation provide a meaningful comparison if baseline data is collected prospectively.

B.3.c: What measures can assess improvements in data completeness, accuracy, and consistency?

Three metrics are particularly well-suited to this evaluation: missing data rate by field and by investigator location, query rate per 1,000 data points (the rate at which entered data requires clarification), and data lock cycle time. These are measurable at baseline without AI and can be tracked through the pilot to isolate AI's contribution. Investigator location-level variation in these metrics is particularly informative: AI tools that reduce average query rates but increase variation across locations may be improving data quality in some contexts while masking problems in others.

B.4 AI System Performance

B.4.a: What metrics are most appropriate for evaluating AI model accuracy, robustness, and generalizability?

Standard classification metrics (sensitivity, specificity, AUC) are appropriate for tools with discrete outputs such as safety signal classifiers or eligibility screeners. For tools producing continuous outputs such as dose recommendations or risk scores, calibration metrics assessing whether predicted probabilities match observed outcomes are more informative than accuracy alone. Robustness should be assessed through performance on held-out subgroups defined prospectively, not post hoc. Generalizability requires testing on data from investigator locations not included in the training set, which needs to be built into study design from the start.

B.4.b: How should AI system stability over time be measured, including model drift?

Model drift is a genuine risk in trials extending 12 to 24 months, where participant population characteristics, standard of care practices, or investigator location behaviors may shift. Drift detection should be built into the deployment architecture from the outset. At minimum, participants should document the frequency of model performance monitoring, the threshold at which retraining or recalibration is triggered, and the process for communicating model updates to trial stakeholders. FDA should define these as minimum documentation standards for pilot participation.

B.4.c: How can performance be evaluated across participant populations, investigator locations, and therapeutic areas?

Subgroup performance analysis should be pre-specified rather than exploratory, with investigator location and participant demographic subgroups defined before deployment. FDA should require pilot participants to report location-stratified performance data, not just aggregate metrics. In our experience deploying AI tools across multiple investigator locations, location-level variation in data quality and clinical practice is one of the most consistent sources of model performance degradation, and it is the variation least likely to be surfaced by aggregate reporting alone.

B.5 Trustworthiness***B.5.a: What evidence should demonstrate that AI systems are valid and reliable?***

The NIST AI RMF treats valid and reliable as the foundational condition on which all other trustworthiness characteristics depend. For AI tools in clinical trials, validity cannot be assessed at the model layer alone. A model that produces valid outputs when given clean, complete, timely data may produce unreliable outputs when that data arrives late or inconsistently formatted. Validity evidence should therefore address both layers: prospective validation on a held-out dataset reflecting the intended use population, documentation of how the model behaves when input data is degraded, delayed, or outside the training distribution, and a clear articulation of known failure modes at both the data pipeline and model levels. Tools that have only been validated in controlled development environments, without exposure to real investigator location data variability, should be required to document that limitation explicitly. We require this of ourselves, as our own tools are validated against development data and then evaluated in deployment.

B.5.b: How should safety and risk mitigation be evaluated?

Against a pre-specified risk register that participants develop before deployment. The register should include, at minimum: risks associated with model failure (false negatives in safety monitoring being the highest priority), risks associated with over-reliance on model outputs by clinical staff, and risks associated with data pipeline failures that degrade input quality without triggering a visible error. Each risk should have a defined mitigation and monitoring approach. FDA should review risk registers at intake rather than treating risk management as a self-certification exercise.

B.5.c: What metrics can assess transparency and explainability for different stakeholders?

Explainability requirements differ by stakeholder in kind, not just degree. An investigator deciding whether to act on a safety alert needs to understand what data drove the alert and whether it is consistent with their direct clinical observation. A data safety monitoring board needs the model's performance characteristics and known failure modes. A regulator reviewing a submission incorporating AI-assisted analysis needs documentation of training data, validation approach, and the accountability chain from model output to human decision.

The pilot should develop a tiered explainability framework that defines minimum disclosure requirements for each stakeholder type, and require that framework to be satisfied as a condition of participation rather than an outcome to be evaluated. A tool that cannot meet minimum explainability standards for investigators, data safety monitoring boards, and regulators before deployment should not be in the pilot. Treating explainability as a metric to be

measured after the fact risks generating evidence about tools that cannot ultimately be used in regulated settings.

On what is technically feasible: from our own experience developing AI tools for deployment in live trials, the information needed to satisfy stakeholder-specific explainability requirements is available at the point of deployment. The gap is not technical; it is that vendors are not currently required to provide it. FDA should define the minimum disclosure requirements for each stakeholder tier before participant selection begins, and use them as an entry criterion. This protects the scientific integrity of the pilot and establishes a standard the field can build toward.

B.5.d: How should privacy protections and data governance practices be evaluated?

Privacy evaluation should assess three points in parallel, as they answer different questions and are often conflated in practice: the technical controls protecting participant data within the AI system; the governance processes determining who can access model inputs, outputs, and reasoning traces; and the auditability and retention posture for AI interactions involving trial data.

Technical controls should be evaluated against applicable standards: HIPAA for US trials and the applicable local data protection regime for international investigator locations, including UK GDPR and the Data Protection Act 2018, EU GDPR, and any country-specific health data law. FDA should consider whether demonstrated alignment with applicable international data protection frameworks can satisfy equivalent privacy requirements rather than requiring duplicative compliance.

Where AI features are used to process trial data, FDA should require participants to evidence contractual guarantees that foundation model vendors cannot train on customer data. Governance processes should be evaluated against a defined accountability structure: a named accountable owner for AI risk, a multidisciplinary review process scaled to risk level, and a written change-control process for foundation model version updates.

Regarding auditability: because foundation model vendors do not generally offer deterministic prompt-to-output reproducibility, the audit log is the only authoritative record of what the AI actually produced. FDA should clarify what logs are required, as AI audit logs may be legally and operationally significant in ways that logs for deterministic code are not. At minimum, FDA should require: an immutable, append-only, tamper-evident audit log for every AI interaction touching trial data; capture of caller identity, timestamp, model identity and version, prompt template version, system prompt, user input, retrieved context, tool calls, final output, and for agentic actions, post-action state and approval IDs; retention policies aligned with the trial's retention period; logical isolation between studies so that logs from one sponsor's trial are not surfaced into AI features serving another sponsor's trial except via a documented data governance gate; and sponsor entitlement to inspect the audit trail for any AI feature touching their study.

Pilot participants should be required to document their data governance practices against these three dimensions as part of the pilot application. FDA should publish minimum standards in advance rather than evaluating them case by case, which will produce more useful comparative data and avoid uncertainty about what will be assessed.

B.5.e: What approaches should assess fairness across demographic and clinical subgroups?

Standard fairness evaluation approaches assume sufficient data in each subgroup to detect performance disparities. Early-phase trials, particularly rare disease studies and first-in-human trials, often enroll participant populations too small to support subgroup-level performance

analysis. Applying conventional fairness metrics in these settings will either produce unreliable estimates or require pooling across trials in ways that obscure meaningful variation.

We recommend a prospective fairness planning approach rather than a retrospective evaluation approach. Before deployment, each AI tool should document the demographic composition of its training data, the subgroups for which performance has been validated, and the subgroups for which performance is uncertain. During the pilot, the focus should be on identifying cases where model behavior appears inconsistent across subgroups, flagging them for human review, and documenting the findings. This produces meaningful learning without overstating what small samples can support.

B.6 Comparative Evaluation

B.6.a: What comparators are most appropriate?

Historical controls are a practical starting point for most pilot metrics, but their validity depends on how well the historical population matches the pilot population. One approach worth considering is matching each AI-enabled trial against a historical twin: a prior study with comparable design, therapeutic area, phase, and endpoint class. This provides a contextually anchored benchmark, but requires tight matching criteria and careful accounting for confounders including changes in standard of care, site experience, and data collection practices over time. FDA should require participants to document the basis for historical comparisons and flag any known differences between historical and pilot cohorts. Concurrent non-AI trials are a more rigorous comparator where they exist, but should not be required as a pilot condition, as this would significantly constrain participant eligibility.

B.6.b: How should differences in trial design, complexity, or therapeutic area be accounted for?

Stratified reporting by trial complexity tier and therapeutic area is the minimum requirement. FDA should work with pilot participants to develop a standardized trial complexity index applied consistently across the pilot, so comparisons can be appropriately adjusted. A small number of pre-specified factors, covering number of dose cohorts, number of investigator locations, participant population complexity, and adaptive design elements, will capture most of the relevant variation.

B.7 Qualitative Outcomes

B.7.a: How can stakeholder trust in AI-enabled trial approaches be assessed?

Through structured interviews and surveys at multiple time points: before AI deployment to establish baseline attitudes, at mid-trial to capture how trust has evolved through early experience, and at trial completion. Surveys should distinguish between trust in the specific AI tool being used and trust in AI in clinical trials more broadly, as these often diverge. For organizations like ours that both develop and deploy the tools, investigator trust is not just a metric; it is operational feedback on whether the tool is calibrated to clinical reality, and it should be tracked accordingly.

B.7.b: What methods can evaluate usability and integration into clinical workflows?

Through direct observation and structured interviews with the clinical staff who interact with AI outputs most frequently: trial coordinators, investigators, and data managers. Key questions include whether AI alerts are interpretable without additional explanation and whether staff

systematically override or ignore AI outputs. Elevated override rates at the investigator location level may point to explainability problems as much as usability problems. If investigators are consistently bypassing AI recommendations, the most likely cause is that outputs are not calibrated to how clinicians reason about the decision at hand. The pilot should track override rates prospectively as a leading indicator.

B.7.c: How should perceived value, scalability, and operational feasibility be measured?

Perceived value should be measured across the full trial rather than against isolated functions. A tool that objectively improves data quality at one stage may not be experienced as valuable by stakeholders who see it primarily generating additional queries elsewhere, and a tool that performs well on a discrete task may still fail to shift the overall trajectory of the trial. The most meaningful measure of perceived value is whether stakeholders believe AI-generated insights at defined intervals during the trial were predictive of where the trial ultimately landed. That framing ties the qualitative assessment directly to the parallel-run mechanism the pilot depends on: AI outputs recorded throughout the trial, compared against final results, with stakeholder perception of value assessed against that same comparison. Scalability should be assessed along two dimensions: technical scalability, covering whether the tool can handle larger data volumes, more investigator locations, and longer trials without performance degradation; and organizational scalability, covering whether sponsors with less mature data infrastructure can realistically deploy and maintain the tool. The pilot will naturally skew toward participants with above-average capability, and FDA's final reporting should be explicit about which findings are likely to generalize to less-resourced settings.

Closing

Five recommendations from our responses warrant particular attention:

- Require two-layer performance reporting that separates data pipeline metrics from model metrics, and analyze integrated versus bolt-on architectures as a primary stratification variable.
- Include in the pilot's governance structure both service providers that execute on behalf of sponsors and organizations that carry both sponsor and CRO accountability for AI deployment, recognizing these as distinct roles with distinct accountability obligations.
- Build the parallel-run evaluation mechanism into the pilot design from the start, so that AI outputs recorded during trials can be compared against final results across multiple studies.
- Assess AI maturity across four independent dimensions at intake, with documented production deployment weighted above claimed modeling capability.
- Require explainability disclosure as a condition of participation, with stakeholder-specific standards defined before participant selection begins.

The pilot has an opportunity to generate not just evidence about AI tools, but evidence about the organizational and infrastructure conditions that determine whether those tools produce reliable output in live trials. That is a more durable contribution than performance data from well-resourced sponsors under favorable conditions. We hope these comments are useful to FDA as the pilot design is finalized and welcome any opportunity to discuss them further.