June 24, 2026

Dr. Mallika Mundkur, Deputy Chief Medical Officer
Office of the Commissioner, Food and Drug Administration
10903 New Hampshire Ave, Silver Spring, MD 20993

RE: ***AI-Enabled Optimization of Early- Phase Clinical Trials Pilot Program; Request for Information***
Docket (FDA-2026-N-4390)

Dear Dr. Mundkur,

Founded in 2001, the Association of Clinical Research Organizations (ACRO) is non-profit trade association representing the world's leading clinical research and technology organizations, which provide specialized services that are integral to the development of drugs, biologics and medical devices that enable patients to live longer, healthier, and more productive lives. ACRO expertise covers the entire spectrum of development – from preclinical, proof of concept, and first in human studies through post-approval, pharmacovigilance, and health data research. ACRO member companies employ nearly 470,000 people worldwide and conduct research in more than 90 countries in every global region.

Regulators around the globe are grappling with the complex challenges of designing regulatory regimes for AI oversight that foster patient safety, data integrity, and high evidentiary standards – without stifling innovation. FDA has a global reputation for scientific expertise and forward leadership that embraces a risk-based, flexible, and pragmatic approach to clinical trial oversight – on topics including artificial intelligence, decentralized trials, risk-based quality management, and integrating RCTs into routine care, to name just a few examples. This RFI (and the broader real-time, continuous clinical trial initiative which it is part of) exemplifies the Agency's global leadership and influence.

Section One – General Comments:

ACRO wishes to acknowledge the scaffolding of standards and guidance within which a successful real-time clinical trials transformation must operate, and we thank FDA for also recognizing this scaffolding in both the RFI announcement and the May 15 Industry Day. These essential standards and guidances include NIST AI Risk Management Framework (AI RMF 1.0), ISO/IEC 42001 and 27001, ICH E6(R3) and ICH E9(R1), and FDA's January 2025 draft guidance on AI for regulatory decision-making.

Before turning to the specific RFI questions posed by the Agency, we wish to emphasize five broader points to facilitate the success of real-time, continuous clinical trials.

I. CROs as a keystone in building the architecture for real-time, continuous clinical trials

CROs are pivotal to the success of the real-time, continuous clinical trials initiative. CROs have dramatically transformed from their origin in the 1980s as arms-and-legs extensions and overflow vendors for their biopharmaceutical clients to their role today as essential strategic partners who catalyze drug development via unique frontline, on-the-ground expertise in the conduct, management, and *innovation* of global clinical trials. CROs today drive success and innovation in a clinical trial ecosystem that is increasingly digitized, personalized, and geographically global. CROs are essential to this initiative's success for five key reasons:

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

1. CRO representation in any RTCT initiative is particularly important for practical implementation and future scale, as CROs act as the integration layer across sponsors, sites, technology providers, and FDA for clinical operations, privacy-preserving analytics, and validated AI governance.

2. Emerging biopharma companies (EBPs) play a fundamental and increasingly large role in modern, global drug development – and CROs play a crucial role for EBPs as essential partners.

   According to the *IQVIA Institute Global R and D Trends 2026:*
   - EBPs accounted for 68 percent of trial starts in 2025
   - EBPs accounted for 86 percent of CAR-T trials, 94 percent of stem cell trials, and 85 percent of non-cellular gene therapy trials over the last five years
   - EBPs originated 61 percent of new drugs in 2025

   Because EBPs must grapple with a narrow financial runway of extreme funding constraints and navigate a complex, dynamic regulatory policy landscape, EBPs rely heavily on their CRO partners.

3. CROs operate the trial in a way that no other party does, and the AI tools most likely to deliver value in this pilot are deployed at the CRO operational layer. ACRO members are currently utilizing AI at every single step of the clinical research process—from project development to study closeout.

4. Small and mid-sized biotechnology sponsors, which are the majority of early-phase IND activity, access AI capabilities through CROs and specialized vendors, so any pilot that recruits only sponsors will systematically under-represent the segment of activity where the pilot's findings would have the greatest practical impact.

5. CROs conduct early-phase trials across dozens of sponsors and hundreds of sites. CROs are positioned to evaluate AI tools across many sponsors, sites, and protocols simultaneously, generating a class of evidence no single sponsor can match. In 2025 alone, ACRO members conducted or participated in 7,634 studies involving 1.4 million patients and volunteers across a wide range of therapeutic areas and geographies that provide drugs, biologics, and medical devices integral to delivering better health outcomes.

Because of CROs' central integration role and their uniquely expansive line of sight across global clinical trials, ACRO asks FDA to explicitly designate CROs as eligible direct participants in the pilot, not as subcontractors visible only through a sponsor relationship.

II. RFI as the beginning of an ongoing, recurring dialogue between FDA and industry stakeholders

We view this RFI as a first step in continuing FDA-industry engagement. The success of both AI-enabled optimization of early phase clinical trials (the topic of this RFI) – and the broader real-time, continuous clinical trial initiative – will be facilitated by ongoing, regular dialogue between the regulator and regulated industry. As FDA has done with other transformative initiatives, such as patient-focused drug development, we urge the Agency to consider creating a suite of recurring multi-stakeholder workshops on real-time, continuous clinical trials. In addition to multi-stakeholder workshops, transparency and engagement would be further supported via a layered or iterative program architecture – with perhaps discrete 3–6-month evaluation sprints nested within an 18–36-month overall program with rolling cohort entry.

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

III. Recommendation for more precision regarding use of the term electronic health records (EHRs)

We ask the Agency to consider more precision and carefully circumscribed use of the term "electronic health records (EHRs)." While FDA as a regulatory authority operates under exemption from the requirements of the Health Insurance Portability and Accountability Act (HIPAA), its industry partners and collaborators do not. FDA discussions and documents within the real-time, continuous clinical trials initiatives, at times, seem to use the terms "EHR" and "EDC" (electronic data capture systems) interchangeably. Because industry stakeholders cannot directly access patient EHRs without explicit patient authorization, we ask the Agency to discontinue the use of the term "EHR' and instead use the term "EDC system." Pilot guidance should reference EDC systems and, where applicable, sponsor-controlled clinical data repositories. This is a material distinction that affects consent design, data governance, and inspectional posture.

In practice, critical trial data already resides across multiple operational systems, including EDC, CTMS, eTMF, safety, and workflow platforms. These systems support active clinical trials and cannot realistically be consolidated into a single centralized environment without introducing significant operational disruption, governance complexity, validation burden, and cybersecurity risk.

Instead, ACRO supports privacy-preserving architectures in which operational signals such as delayed safety review, enrollment slowdowns, monitoring delays, overdue training, or protocol deviation trends are identified within the systems where trial execution is already occurring. Structured alerts, summaries, and risk indicators can then be shared centrally for oversight without requiring wholesale movement of underlying operational or patient-level data.

RTCT infrastructure should also support movement toward structured, machine-readable regulatory information exchange. While early pilots may focus on operational signals, longer-term success will depend on reducing reliance on static, document-heavy workflows and enabling structured data and metadata exchange that can be reviewed by both humans and automated systems. ACRO believes this approach is more scalable, operationally feasible, and aligned with modern privacy and cybersecurity expectations than centralized aggregation architectures.

IV. Clinical Trial Modernization Act

The RFI Federal Register announcement rightly highlights the challenges early trials face [pg. 2301] including facing limited patient populations. Improving access to clinical trials amongst underrepresented populations requires reducing barriers to participation. This will then increase the reach of the pilot program and encourage participant enrollment in the program as a whole.

The pilot program can ensure representation amongst participants by adopting initiatives to encourage participation amongst underrepresented groups and reducing barriers to participation. Opportunities to do so are included in H.R. 3521 – Clinical Trial Modernization Act. The Act encourages the provision of free digital health technologies that are necessary to facilitate trial participation amongst underrepresented patient populations and encourages cost-sharing and other means of financial support for clinical trial expenses to increase participation amongst these groups. Adopting such provisions, amongst other such opportunities to diversify patient populations, will broaden the applicability of the pilot program's outcomes.

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

V. FDA FY 2027 Performance Budget Overview

ACRO supports FDA's legislative proposal, included in FDA's Fiscal Year 2027 Budget Request, to accelerate US-based early phase clinical trials – *"Create a New Clinical Trial Notification Pathway to Serve as an Alternative to the Burdensome Existing Investigational New Drug Pathway to accelerate drug development timeline to Make America Healthy Again"* – which would help facilitate the objectives of the RFI pilot project. The advancement of Phase I research in the US will benefit from both this RFI and FDA's legislative proposal.

Section Two – ACRO Responses to RFI Questions:

RFI Section A: Pilot Program Design and Implementation

We support structuring the pilot to maximize learning and impact by prioritizing use cases, participants, and evaluation approaches that reflect the operational realities of clinical trial execution. To drive meaningful operational improvements, the pilot frameworks should focus on workflow integration and adaptive collaboration across sponsors, sites, CROs, and FDA.

1. Scope and Focus

1(a) Which trial types or trial issues might benefit most from the application of AI (e.g., first-in-human, oncology dose escalation, rare disease trials)?

AI provides the greatest impact in complex, high-risk, and data-intensive trial types—such as oncology, rare disease, early-phase, and multi-site studies—where predictive insights and proactive monitoring can significantly improve safety, efficiency, and outcomes. In addition, the pilot should explicitly include pediatric and other small-N studies and cell and gene therapy programs that require intensive safety surveillance, where the decision density per patient is highest. Evaluation should also recognize that many early-phase delays are operational rather than algorithmic, so AI-enabled approaches should be assessed in parallel with complementary non-AI operational improvements (e.g., streamlined startup, centralized monitoring under ICH E6(R3), and RTCT-enabled data flows).

The pilot should prioritize the interoperability, workflow integration, and governance models required to bridge fragmented infrastructure, which remains the primary barrier to scalable AI-enabled trials. While it may be beneficial to focus initial pilot activities on digitally mature organizations, the broader impact will depend on improving research site enablement and recruitment workflows across environments with varying levels of operational maturity. Focusing exclusively on resource-rich research sites risks validating AI performance only within optimal conditions; therefore, it is critical for follow-on phases of the pilot program to include community-based and independent research sites to ensure findings are generalizable across the entire clinical trial ecosystem.

Priority areas should include data review, query management, safety monitoring, protocol deviation detection, and interoperability across clinical and regulatory systems. Specifically, AI can address significant manual burdens, such as the recreation of schedule-of-events information across EDC, CTMS, and site workflows, or the reconciliation of adverse event data across EHR, EDC, safety, and regulatory systems.

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

Complexity-driven studies—including first-in-human, oncology dose escalation, and rare disease trials—stand to benefit most due to their inherent complexity and the need for rapid decision-making. Ultimately, the broader value of the pilot will come from demonstrating that these AI-enabled operational improvements are scalable and reproducible across diverse clinical trial settings.

1(b). Should the pilot target specific therapeutic areas or remain broadly applicable?
The pilot should remain broadly applicable across diverse therapeutic areas, patient populations, and research settings to evaluate AI performance across the full clinical trial ecosystem. Ensuring broad applicability will better inform the development of future standards, guidance, and evaluation frameworks for AI-enabled clinical research technologies. To enable meaningful cross-participant comparison while preserving broad applicability, each pilot participant should designate a single primary use case with pre-specified endpoints, a stated risk tier, and proportional governance controls.

1(c) Should priority be given to specific AI use cases (e.g., recruitment, safety monitoring)? If so, which?
We believe both recruitment and safety monitoring would be worthy use cases. We recommend prioritizing AI use cases that address well-defined operational challenges and integrate into existing clinical workflows. Priority areas should include participant identification and eligibility screening, safety signal detection, protocol deviation review, data quality review, and query management. FDA should focus on use cases where AI augments – rather than replaces – decision-making, particularly in regulated workflows requiring traceability and oversight. Notably, three of these priority use cases are operationally located at the CRO, which is a pivotal reason that CRO participation cannot be optional in pilot design.

Priority use cases should also include dose-escalation and expansion decision support (with mandatory human-in-the-loop and pre-specified escalation rules) and contextual evidence integration to inform protocol and cohort decisions. For an inaugural cohort, oncology safety surveillance offers a feasible path to validate the foundational architecture – data ingestion, transmission integrity, blinding preservation, and auditability – before expanding to higher-risk statistical automation.

2. Participant Selection

2(a). What criteria should FDA use to select sponsors, trials, or technologies?
Sponsors should be a representative sample of 3 segments: emerging biopharma, mid-size, and large biopharma. FDA should prioritize sponsors, trials, and technologies that enable measurable evaluation of AI within clinical trial operations. Selection criteria should focus on assessing operational impact, workflow integration, reproducibility, and performance across diverse research settings. Priority should be given to technologies already in active clinical use and capable of flexible deployment across varied operational environments. Final selection should ensure representation across a vast range of therapeutic areas, including rare diseases and high-unmet-need populations, ensuring the pilot's findings are broadly generalizable across the full spectrum of clinical research.

Selection criteria should also stratify across geographic and site mix, trial complexity (first-in-human vs. later Phase 1b), and solution maturity (assistive analytics vs. higher-automation workflows). Each selected participant should submit a credible validation plan and a concise system or model card covering intended use, data lineage, performance claims, monitoring, and rollback procedures, aligned with recognized AI risk-management and management-system standards. Selection criteria should also include key elements outlined in the FDA 2025 AI guidance – including a description of the AI model's context of use, its influence

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

on the operational or regulatory decision, the consequence of model failure, and the performance thresholds the model must meet before deployment.

2(b). How can the pilot ensure representation across organization size, capability, and therapeutic areas?
The pilot should include a diverse range of participants, including large and small sponsors, academic medical centers (AMCs), site networks, and community-based research settings. While AMCs remain vital to clinical innovation, expanding participation to include independent and community-based sites is critical for improving patient access and supporting scalable trial execution. For pilot inclusion, FDA should consider collaborating with AMCs that have long-standing relationships with community-based sites, utilizing a hub-and-spoke model, to ensure that the AI technology can scale to include a broader range of organizations and participants. Representation across diverse therapeutic areas, patient populations, trial phases, and geographic regions will help ensure AI-enabled approaches are broadly applicable and support continued operational leadership throughout the clinical trial ecosystem. Representation can be further strengthened through reserved pilot slots for smaller and emerging innovators, optional pairing with an experienced clinical operations partner, and rotational entry across participant archetypes.

3. Collaboration Models

3(a). What partnerships (e.g., sponsor-tech vendor-academic-FDA) are most effective?

Effective partnership models should include sponsors, research sites, technology providers, standards organizations, trade associations, CROs, patient organizations, and FDA. Because AI-enabled approaches impact multiple stages of execution, evaluations should reflect collaboration across recruitment, data collection, safety review, and regulatory workflows. Initial partnerships should focus on these core stakeholders to simplify implementation and operational coordination during early proof-of-concept activities. To maximize learning, FDA should evaluate different combinations of sponsors, sites, and technology providers to identify which collaborative models drive the greatest operational consistency. For example, evaluating a single technology provider's performance across multiple sponsor and site environments can reveal whether AI-enabled workflows are truly interoperable or if they require site-specific customization. As pilot activities scale, broader ecosystem participation—including CROs—will be essential to evaluating whether AI-enabled approaches support operational consistency across various research environments.

In these partnerships, FDA adds the greatest value as a neutral convener – providing feedback on evaluation metrics, risk tiering, and RTCT expectations – rather than as an operator of proprietary models, while sponsors retain accountability for investigational-product decisions.

3(b). How can FDA facilitate pre-competitive collaboration and knowledge sharing?
FDA can facilitate pre-competitive collaboration through shared evaluation frameworks, clear pilot documentation, and operational knowledge-sharing mechanisms. The pilot should also promote open standards and interoperability while maintaining robust protections for confidential information, intellectual property, patient privacy, and cybersecurity. Leveraging existing trade groups and standards organizations can further support these efforts by providing established forums for multi-stakeholder dialogue and alignment on AI-enabled methodologies.

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

Pre-competitive artifacts should include reference evaluation scenarios and shared benchmarking taxonomies (e.g., assistive vs. autonomous AI, human-in-the-loop requirements, agent-based systems), along with anonymized lessons-learned memoranda.

3(c). What role should patient groups and investigators play in AI governance?
Patient groups and investigators are essential partners in AI governance, ensuring that AI is designed, validated, meets requirements and used in a way that is patient-centered, clinically appropriate, and ethically sound. ICFs should be designed to include considerations for AI and Governance. Patient groups and investigators are essential for evaluating whether AI-enabled approaches improve the practical experience of trial participation and execution. Their input is critical to identifying usability concerns, operational burdens, and trust considerations that technical performance metrics alone may not capture. To move beyond high-level engagement, FDA should prioritize identifying the specific sponsors and therapeutic areas for each pilot project and then systematically incorporating representative patient groups from those respective areas into the governance process. This targeted participation will help the agency assess how specific AI-enabled approaches affect patient access, site workflows, and overall confidence in the clinical trial process across diverse communities. Governance structures should also provide for continuous participation across the pilot lifecycle, including formal review of model drift, incidents, and changes that could affect participant risk, with investigators helping to define interpretable and actionable workflows.

4. Operational Structure

4(a). What support (e.g., regulatory engagement, technical guidance) should FDA provide?
The FDA should provide clear, risk-based guidance, lifecycle recommendations, and collaborative regulatory pathways to enable safe, transparent, and scalable AI adoption in healthcare while supporting continued innovation.

FDA should provide clear guidance regarding evaluation expectations, documentation, auditability, human oversight, and performance assessment for AI-enabled clinical trial approaches. Shared evaluation frameworks and standards-based approaches will help support consistent implementation and comparability across pilot participants. Specifically, technical guidance should detail how sponsors can submit "platform-level" validation data that can be referenced across multiple trials, moving away from the current trial-by-trial evaluation model.

FDA should also provide timely clarification on how AI-supported outputs may appear in monitoring reports, protocol amendments, and regulatory submissions without creating unintended endpoint or labeling ambiguity.

In addition, we ask FDA to provide worked examples of the January 2025 AI draft guidance credibility framework applied to the specific contexts of use the pilot will evaluate. The draft guidance is general-purpose, and operational application benefits substantially from worked examples.

FDA should also consider structured engagement checkpoints with participants to clarify regulatory expectations and identify implementation challenges. These checkpoints will help participants understand how FDA evaluates workflow integration, data quality, traceability, and operational performance during the pilot.

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

The pilot should utilize and iterate upon secure environments for structured, machine-readable signal transmission via API endpoints, ensuring tech partners can submit metadata summaries directly to the agency's centralized system.

4(b). What infrastructure is needed (e.g., secure data environments, shared tools)?
Infrastructure for AI-enabled trials must prioritize secure and auditable environments for the exchange and review of clinical trial information among all stakeholders. Foundational to this infrastructure is the adoption of standards-based clinical research technologies at the site level—such as direct data capture and digitally enabled workflows—to ensure interoperability across the ecosystem.

Required infrastructure should also include immutable audit trails, versioned datasets and model releases, and interoperable logging of the human decisions taken in response to AI outputs, so that human oversight is itself auditable.

ACRO explicitly cautions against any infrastructure design that requires participants to migrate operational data to a central environment. Operational trial data (e.g., EDC, CTMS, eTMF, safety database) lives in production systems that cannot be relocated for the convenience of a pilot. A federated model in which evaluation metrics are computed locally and aggregated centrally is more realistic.

As part of the pilot, FDA should consider establishing a secure proof-of-concept environment for structured, machine-readable submissions. This environment would help both humans and automated systems identify relevant clinical information within complex submission packages, supporting more scalable regulatory oversight and AI-enabled workflow evaluation.

FDA should consider a federated learning infrastructure to allow sharing of data while keeping privacy and confidentiality agreements in compliance.

4(c). How can the pilot accommodate varying levels of AI maturity across participants?
The pilot should evaluate AI performance across varying levels of digital maturity. Supporting both native AI capabilities and integrations with external services will allow FDA to assess scalability and reproducibility across diverse technology environments.

The pilot could accommodate varying AI maturity through a tiered, modular, and governed approach—enabling all participants to engage safely while progressing toward more advanced AI capabilities over time. In a tiered model, participants are onboarded into foundational, intermediate, and advanced tracks, allowing teams with limited AI experience to start with guided, low-complexity use cases, while more mature organizations engage with advanced analytics and agentic workflows. Higher-risk tiers (uses affecting dose, eligibility, or regulatory-facing conclusions) should require stronger pre-specified validation, monitoring, and independent checks.

5. Timeline and Milestones

5(a). What is an appropriate duration for the pilot?
We recommend an initial pilot period of approximately 12 months to support participant onboarding, workflow stabilization, and early operational evaluation. While certain workflows—such as structured regulatory submission and review—may demonstrate measurable benefits quickly, complex activities involving patient recruitment, protocol execution, and longitudinal trial operations require longer evaluation

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

periods to assess reproducibility and scalability. A phased pilot structure with interim evaluations will support iterative learning and help FDA and participants identify scalable operational practices over time.

This initial period should be understood as the first evaluation window within a longer-running learning agenda. A layered structure – shorter 3–6-month evaluation sprints nested within an approximately 18–36-month overall program with rolling cohort entry – would allow the pilot to capture both near-term operational wins and longer-horizon evidence on decision quality, model drift, subgroup performance, and generalizability.

5(b). What interim milestones or checkpoints should be included (e.g., enrollment, safety review, interim analyses)?

Interim milestones should include measurable operational and regulatory outcomes relevant to clinical trial execution. In addition to traditional milestones like enrollment and safety review, FDA should prioritize tracking data accuracy, reductions in manual processes, duplicate data entry, query cycle times, and protocol deviation review timelines.

Where applicable, interim milestones should also include dose-escalation and expansion decisions and – for early cohorts – demonstrated feasibility and quality of real-time safety monitoring, with each evaluation sprint producing a pre-specified deliverable package covering methods, metrics, incidents, and human-oversight records.

Milestones should also evaluate whether AI improves the consistency and usability of regulatory submissions. For example, structured and machine-readable workflows can be assessed by their ability to reduce duplicative documentation and streamline review processes for both humans and automated systems. Additional checkpoints should assess reproducibility across diverse operational settings, the auditability of AI-supported processes, and the maintenance of appropriate oversight throughout the trial. These ongoing checkpoints will facilitate shared learning among participants and help identify where AI-enabled approaches demonstrate meaningful improvements in scalability.

5(c). How should FDA balance rapid insights with rigorous evaluation?

The FDA should adopt a risk-based, lifecycle approach that enables rapid generation and application of insights while maintaining rigorous validation for higher-risk use cases—supporting iterative testing, real-world evidence, and continuous monitoring, with stricter controls applied where patient safety or clinical decision-making is impacted.

Most importantly, sprint-level outcomes from program-level conclusions. Rigor can be strengthened through pre-registration of claims and metrics, methods-forward interim learning reports that document failures as well as successes, and independent review for higher-risk tiers – with sprint-level outcomes kept distinct from program-level conclusions so near-term progress is not conflated with longer-horizon validation.

FDA should differentiate between operational capabilities that provide clear, reproducible improvements—such as reductions in manual processes and documentation—and longer-term activities like recruitment and protocol execution that require extended evaluation.

Participants should establish governance frameworks and procedures documenting how AI-enabled approaches are evaluated and overseen within clinical workflows. Periodic reporting on progress, challenges,

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

and lessons learned will support shared learning while ensuring the oversight, traceability, and reproducibility necessary to prove that improvements are reliable, scalable, and aligned with regulatory expectations.

6. Knowledge Sharing

6(a). How should lessons learned be captured and disseminated?

Pilot participants should file a report at the end of their project – to be shared in an aggregated, de-identified manner via FDA summary reports. FDA should implement mechanisms for periodically sharing aggregated, de-identified insights—including reproducibility findings and workflow considerations—via structured pilot reports, workshops, or collaborative forums. These sharing models must maintain robust protections for confidential information, intellectual property, patient privacy, and cybersecurity. Ultimately, consistent, structured knowledge sharing should help inform future guidance development, drive standards-based approaches, and support broader modernization efforts across the clinical trial ecosystem. To improve cross-participant comparability, FDA could also adopt standardized incident and drift reporting templates and structured after-action reviews as part of the knowledge-sharing toolkit.

6(b). What mechanisms can promote transparency while protecting proprietary information?

Transparency can be achieved while protecting proprietary information through a combination of controlled disclosure, technical safeguards, and governance mechanisms. This includes providing auditability, explainability of outputs, and traceability of data sources without exposing underlying model IP, alongside clear documentation of model purpose, limitations, and performance characteristics. Approaches such as federated learning infrastructure further support this balance by enabling model training and validation across distributed datasets without sharing or centralizing sensitive data, preserving data ownership and confidentiality. Additional mechanisms include secure enclaves, access controls, model cards, and standardized reporting frameworks, ensuring regulators and stakeholders have visibility into model behavior and risk controls while maintaining protection of proprietary algorithms and data assets.

Knowledge-sharing mechanisms should balance broader operational learning with the protection of intellectual property, patient privacy, and cybersecurity. Shared learning is most effective when focused on evaluation methodologies, operational findings, and standards-based practices rather than the disclosure of proprietary algorithms, models, or sensitive clinical data.

FDA should consider collaborative frameworks and trusted industry initiatives to support neutral evaluation environments and standards development. Additionally, secure and neutral submission and review environments can facilitate structured regulatory oversight and auditability while allowing participants to maintain governance over sensitive technologies. Common documentation frameworks and aggregated operational summaries will provide necessary visibility into pilot outcomes without compromising proprietary information.

RFI Section B: Evaluation Metrics and Success Criteria

The pilot should evaluate whether AI-enabled approaches produce measurable operational improvements across clinical trial execution. Evaluation metrics should assess both overall development outcomes and the operational workflows contributing to trial efficiency, quality, and scalability.

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

1. Trial Efficiency and Speed

1(a). How should improvements in trial efficiency be measured (e.g., time to initiation, enrollment, or completion)?
Improvements in trial efficiency should be measured using both high-level development milestones and operational metrics reflecting how work is performed across the trial lifecycle. Beyond time to initiation, enrollment, and completion, FDA should track reductions in manual processes, duplicate documentation, query cycle times, and protocol deviation or safety review timelines. FDA should also consider metrics related to workflow integration, interoperability, and the adoption of structured, machine-readable submissions across sponsors, and sites. Evaluating these operational metrics ensures that observed improvements are driven by meaningful operational changes—rather than isolated gains—that are reproducible and scalable across diverse research settings. Where RTCT or AI-enabled workflows are new, initial measures of success may appropriately focus on feasibility and quality of real-time processes (for example, safety-monitoring integrity) rather than assuming immediate calendar compression of end-to-end trial timelines.

1(b). What metrics should be used to assess reductions in time from Phase 1 completion to Phase 2 initiation?
Measuring reductions should focus on end-to-end cycle time from Phase 1 completion to Phase 2 initiation, supported by milestone-level timing, decision speed, and efficiency gains across study start-up and operational readiness.

FDA could require a single pre-specified primary interval per entrant (for example, calendar time from a pre-specified Phase 1 primary completion or interim analysis to Phase 2 protocol finalization and first Phase 2 site initiation visit). Parallel 'at-risk' Phase 2 planning may be reported where supported by interim integrated safety, PK, and PD review, with explicit gates and documented rollback criteria.

Metrics evaluating the time from Phase 1 completion to Phase 2 initiation should prioritize the discrete operational activities that often stall development. FDA should consider timelines for protocol finalization, schedule-of-events (SoE) development, site activation, patient recruitment preparation, and regulatory submission and review activities. While these metrics provide a baseline for evaluating AI-enabled coordination, the pilot must recognize that phase transitions are not operationally equal. The transition from Phase 1 to Phase 2 typically involves a limited number of sites and highly controlled populations. Therefore, while these results are foundational, they cannot be extrapolated as a one-to-one proxy for later transitions. Instead, they should be used to assess how AI handles the initial shift in complexity, with the understanding that Phase 3 transitions require managing a geometric increase in site diversity, global regulatory volume, and patient enrollment scale.

1(c). How can improvements in participant screening, recruitment efficiency, and participant retention be quantified?
Improvements should be quantified using traditional enrollment metrics paired with operational measures reflecting workflow sustainability. FDA should consider time required to screen eligible participants, consistency of eligibility assessments, enrollment timelines, dropout rates, and protocol adherence across diverse research settings. Crucially, evaluation must assess whether AI reduces operational burdens such as manual screening activities, fragmented data review, and repetitive documentation across both sponsors and research sites.

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

Metrics should also include participant engagement, usability of consent and communication workflows, and population diversity across community-based and underserved research settings. Where feasible, comparative assessments of AI-enabled versus non-AI workflows within the same study or therapeutic area—stratified across academic medical centers, site networks, and community organizations—will help identify whether observed improvements are truly attributable to AI-enabled operational workflows rather than variations in site infrastructure, patient populations, or organizational models.

2. Decision Quality

2(a) How can the quality and timeliness of go/no-go decisions be evaluated (both FDA regulatory decisions and sponsor-internal decision points)?

The quality and timeliness of go/no-go decisions can be evaluated through a combination of decision speed, evidence quality, and outcome alignment, applied consistently across both FDA and sponsor processes. Key metrics include time to decision from data availability, completeness and robustness of supporting evidence (e.g., data quality, statistical confidence, and risk assessment), and decision consistency with predefined criteria or regulatory guidance. Effectiveness is further assessed by downstream outcomes, such as reduced rework, fewer protocol amendments, improved study success rates, and alignment with safety and efficacy expectations—ensuring decisions are both fast and well-founded.

ACRO emphasizes two key considerations here:

- The quality and timeliness of go/no-go decisions should be evaluated based on decision outcomes and the fidelity of the operational processes supporting them. Central to this evaluation is the preservation of the sponsor's role as the "Clinical Fiduciary." While AI-enabled platforms can automate the aggregation of disparate data and identify statistical anomalies, they cannot replace the sponsor's essential responsibility for contextualizing safety signals and providing medical oversight.

- The principal risk is the conflation of decision speed with decision quality: a faster decision that proves wrong is not an improvement. We recommend that the pilot prospectively document, at each go/no-go decision: the evidence considered, the model output considered, the human reviewer's assessment, the decision made, and the reasoning. Without this prospective documentation, decision quality cannot be evaluated at all, retrospectively or otherwise.

Evaluation should therefore include specific operational measures that empower sponsor-led decision-making, including:

- Reduction of "Administrative White Space": Measuring reductions in delays associated with data aggregation, data cleaning, and manual safety review workflows, thereby minimizing the administrative dead time between trial phases.
- Enhanced Decision Fidelity: Assessing whether AI tools allow sponsors to more effectively synthesize protocol interpretations and set medical thresholds (e.g., Dose Limiting Toxicities) based on real-time data.
- Auditability and Reproducibility: Ensuring that underlying workflows remain transparent and that AI-supported decisions are well-characterized and reproducible across participating organizations.

Ultimately, maintaining rigorous human oversight ensures that AI-enabled decision support remains clinically sound. Success should be measured by how effectively these tools improve the consistency and scalability of a sponsor's judgment rather than simply accelerating isolated processes.

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

2(b) What methods can assess concordance between AI-supported and traditional decision-making?
Concordance between AI-supported and traditional decision-making can be assessed by side-by-side comparison of AI outputs against subject-matter expert judgment and existing standard processes, documented in validation evidence (test outputs, data versions, and parameter settings) and reviewed through governance/QA sign-off; if AI performs worse or is not clinically reasonable, it does not progress beyond pilot.

Concordance should be assessed through comparative evaluation where AI-enabled workflows are analyzed alongside traditional processes in similar clinical environments. These assessments should compare decision outcomes, supporting rationale, workflow efficiency, and consistency of interpretation across participating organizations.

Where feasible, concordance should also be assessed through blinded comparisons on historical or holdout datasets, with pre-specified adjudication when AI-supported and traditional assessments diverge and measurement of downstream rework. Where Bayesian or adaptive frameworks are used, evaluation should include whether prior or predictive-distribution updates remain within pre-specified operating characteristics consistent with ICH E9(R1) estimand principles. For RTCTs where FDA may make earlier decisions prior to study completion, the pilot should also test mechanisms for third-party assessment and validation of signals at scale.

Where operationally feasible, sponsors should evaluate AI and non-AI workflows across comparable research sites—including academic medical centers, site networks, and community-based organizations—within the same study or therapeutic area. This comparative approach helps identify whether improvements are truly attributable to AI-enabled operational workflows rather than differences in site infrastructure, patient populations, or organizational models. Finally, evaluation must prioritize transparency and traceability, ensuring the ability to audit inputs, assumptions, and oversight activities to identify where AI improves consistency and where additional governance is required.

2(c) How should reductions in late-stage trial failures attributable to improved early-phase decisions be measured?
Reduced late-stage failures should be measured by improved Phase II/III success rates and fewer downstream issues, with clear traceability to enhanced early-phase decision quality enabled by AI-driven insights and governance.

While reducing late-stage failures is a critical long-term goal, direct attribution is challenging due to the myriad of factors influencing Phase 3 outcomes. FDA should therefore avoid relying solely on late-stage results as a measure of early-phase decision quality. Instead, evaluation should prioritize leading operational and clinical indicators that more directly reflect workflow improvements.

These include consistency in dose selection, alignment between early safety/efficacy signals and later findings, and reductions in avoidable protocol amendments. Prioritizing these process-oriented measures—focused on the quality, completeness, and timeliness of information—provides more immediate, actionable insight into whether AI-enabled approaches improve the consistency and reproducibility of decision-making across diverse research environments.

3. Participant Safety and Data Integrity

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

3(a) What metrics should be used to evaluate the detection and response time for safety signals?
Evaluation should look at both missed safety signals and excessive false alerts, since both can create operational and patient safety issues. ACRO proposes a decomposed timing measurement (event → entry → coding → review → action). Detection and response time for safety signals should be evaluated through timeliness and effectiveness metrics supported by AI enabled anomaly detection. This includes measuring time-to-detection (e.g., how quickly anomalies or emerging safety signals are identified from data trends), time-to-action (e.g., speed of investigator or study team response following alert generation), and signal confirmation rates (accuracy of detected anomalies requiring intervention). AI tools are unlikely to materially affect the event-to-entry interval (which is a clinical workflow problem) but should plausibly affect the entry-to-coding and coding-to-review intervals. It is useful to set reasonable expectations for what the pilot can attribute to AI.

Safety-signal metrics should also capture alert burden per patient-time at risk and the positive predictive value of prioritized safety reviews, drawing on established pharmacovigilance AI frameworks adapted to interventional early-phase trial governance.

AI-driven anomaly detection approaches—such as trend analysis, outlier detection, and risk-based monitoring—enable earlier identification of deviations or adverse event patterns, which can be assessed by comparing detection timing and intervention outcomes against traditional monitoring benchmarks.

FDA should evaluate both the timeliness and reliability of safety signal detection and response. Key metrics should evaluate the velocity of data flow from initial capture to signal transmission via the API, targeting the agency's specified 24-hour window, paired with metrics tracking the subsequent time to human clinical review and resolution.

Beyond elapsed time, assessment must include the timeliness, completeness, and consistency of safety data flow across systems. Because delays in how AE data is captured, structured, and transmitted significantly impact detection timelines (regardless of analytical power), the pilot should prioritize improving data capture and submission infrastructure. FDA should also evaluate whether structured, interoperable, and machine-readable approaches to safety reporting reduce duplicate reporting and support scalable review across sponsors, sites, and CROs. Ultimately, while AI may improve signal prioritization, its effectiveness depends on underlying data quality; in many cases, improvements to data capture infrastructure may yield greater gains in safety than downstream analytical enhancements alone. Maintaining rigorous human oversight remains critical to ensuring accurate interpretation and response.

3(b) How should the impact of AI on adverse event rates or protocol deviations be assessed?
Assess impact by using a pre/post (or matched control) study design that compares adverse event (AE) rates and protocol deviation rates while holding key confounders constant (study phase, therapeutic area, site mix), supported by a traceability matrix and validation evidence linking AI requirements → tests → outcomes. Timeliness should be assessed by tracking time-to-detection and time-to-action on emerging AE or deviation trends (e.g., earlier signal identification and faster mitigation), and results should be reviewed in quarterly performance reviews with documented deviations and corrective actions when acceptance criteria are not met.

Assessment of AI's impact on adverse event (AE) rates and protocol deviations must account for clinical and operational variables—including patient populations, study design, and site performance—that may

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

influence outcomes independently of AI. To ensure direct attribution, FDA should prioritize process-oriented measures such as earlier detection of safety issues, reduced time to identify and address deviations, improved monitoring consistency, and a reduction in duplicate or incomplete reporting.

Evaluation should also utilize comparative assessments across similar research settings to determine whether observed improvements are attributable to AI-enabled workflows rather than differences in site infrastructure or operational maturity. Ultimately, quantifying improvements in data completeness, accuracy, and timeliness will ensure that changes in AE rates or protocol deviations are supported by reliable, comparable operational data.

3(c) What measures can assess improvements in data completeness, accuracy, and consistency?
Improvements in data completeness, accuracy, and consistency should be measured through data quality and integrity metrics, including percentage of complete data fields, error rates or discrepancy rates, and consistency across sources and time points. These should be supported by validation and audit checks (e.g., reconciliation rates, query resolution times, and traceability of data changes), ensuring that data is accurate, standardized, and reliable for clinical use. Continuous monitoring and periodic review further confirm sustained improvements in data quality and regulatory readiness.

Improvements should be assessed using metrics that reflect both data quality and the operational processes used to generate and manage data throughout the lifecycle. Relevant measures include rates of missing or incomplete data, frequency and resolution time of data queries, cross-system consistency, alignment with standardized data models, and timeliness of data availability for regulatory review. Useful data-integrity indicators also include the rate of late-discovered eligibility issues, which are a frequent source of important protocol deviations in early-phase trials.

FDA should also prioritize metrics related to data provenance, traceability, and interoperability, specifically the ability to track data from source through transformation, aggregation, and submission. While FDA oversight traditionally focuses on data once it enters sponsor-controlled systems, sustained improvements in quality depend heavily on the adoption of standardized, interoperable structures upstream in the research ecosystem. AI-enabled approaches can accelerate review by flagging anomalies, but meaningful gains require well-defined governance and standardized data structures across sponsors, sites, CROs, and technology providers.

4. AI System Performance
4(a) What metrics are most appropriate for evaluating AI model accuracy, robustness, and generalizability?
Paramount Consideration:
The paramount consideration when developing metrics to evaluate AI system performance is that performance metrics must be specified in the context of use. A patient pre-screening model and a safety triage model and a dose-decision support model have different operationally relevant metrics, and a single "AI performance dashboard" that uses the same metrics for all of them will not be informative. Metrics are not universal, but must be pre-specified for a particular context of use.

Additional Considerations:
In addition to context of use, appropriate metrics should align to three dimensions—accuracy, robustness, and generalizability—and be defined within the model validation plan with clear acceptance criteria and auditability.

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

Accuracy:
Standard performance metrics such as precision, recall, F1 score, and error rates, alongside domain-specific measures (e.g., correctness of outputs, completeness of extracted information)

Robustness:
Evaluation under stress and edge-case scenarios, including sensitivity to data variations, stability across inputs, and drift detection over time

Generalizability:
Performance across independent validation datasets, different study contexts, and varied populations, ensuring consistent results beyond the training environment

Supporting metrics:
Reproducibility, explainability, fairness, and audit trail completeness to ensure models are reliable, transparent, and fit for regulated use

Evaluation of AI system performance should consider whether AI-enabled approaches produce reliable, reproducible, precise, and operationally meaningful results across clinical trial environments. Performance assessment should include consistency of outputs across varying data conditions, operational settings, workflows, and levels of data completeness and standardization.

To support technical rigor, evaluation should utilize standard quantitative performance metrics appropriate for specific AI tasks—such as sensitivity, specificity, and Area Under the Receiver Operating Characteristic (AUROC) for classification models, or Mean Absolute Error (MAE) for predictive analytics—stratified across operational settings to assess generalizability. These technical metrics should be paired with "stress testing" against edge-case data to evaluate model robustness in the face of missing or inconsistent source data.

FDA should also consider system-level performance, including whether AI-enabled approaches integrate effectively into existing clinical trial workflows, support traceability and oversight, and improve operational efficiency without introducing unnecessary complexity. Finally, evaluation should account for the quality and consistency of underlying data environments, as variability in data structure and interoperability can significantly influence performance. Assessments should emphasize operational reproducibility across diverse research environments rather than isolated model behavior under controlled conditions.

Predefined thresholds or triggers for when a model's performance degrades and requires review, retraining, rollback, or additional oversight.

4(b) How should AI system stability over time be measured, including detection and mitigation of model drift?

AI system stability should be measured through continuous performance monitoring over time, including tracking accuracy and key outcome metrics against baseline benchmarks, alongside drift detection indicators (e.g., changes in data distributions or model outputs). When drift or degradation is identified, defined thresholds trigger investigation, re-validation, and corrective actions (e.g., retraining or recalibration), supported by documented change control and periodic review cycles to ensure sustained reliability and compliance.

AI system stability should be evaluated through continuous monitoring of operational performance, data quality, and workflow consistency across changing clinical trial environments. Evaluation frameworks must consider how shifts in data sources, study populations, site capabilities, and data collection practices influence system behavior over time.

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

FDA should encourage governance that supports robust documentation of system updates, evaluation methodologies, and oversight throughout the pilot lifecycle. Monitoring should assess not only model drift but also changes in upstream data quality, interoperability, and workflow integration that may impact downstream performance.

Because the frequent introduction of new model versions can complicate longitudinal evaluation and reproducibility, FDA should promote clear version documentation and change management practices. Maintaining high visibility into these methodologies and governance processes across pilot participants will support operational reproducibility, shared learning, and broader ecosystem confidence. Ultimately, prioritizing transparency, traceability, and oversight ensures that AI-enabled approaches continue to perform reliably across diverse operational settings and therapeutic areas.

4(c) How can performance be evaluated across different patient populations, trial sites, and therapeutic areas?

Performance across different patient populations, trial sites, and therapeutic areas should be evaluated through stratified and subgroup analysis, using consistent metrics applied across sites, geographies, and populations to identify variability in outcomes. This includes validating models on diverse datasets spanning multiple studies, sites, and therapeutic areas, ensuring data sufficiency and representativeness while controlling for key variables such as study phase and population characteristics. Results should be monitored through standardized dashboards and site-level performance indicators to ensure consistent performance and detect any population-specific bias or degradation.

Performance should be evaluated across diverse patient populations, trial sites, therapeutic areas, and technology ecosystems representative of clinical research. Evaluation approaches must assess consistency across varying site types, workflow models, and operational maturity levels to determine whether observed improvements are reproducible and scalable across the clinical trial ecosystem.

Evaluation should also include explicit declaration of limits of use, so that the AI-enabled approach is not relied upon where performance degrades outside the populations, sites, or time windows represented in its training data.

Comparative evaluation across similar operational settings will help identify whether observed improvements are attributable to AI-enabled workflows rather than differences in site infrastructure or organizational maturity. Furthermore, evaluation should reflect the reality of the fragmented research data landscape. Because trial execution depends on multiple operational systems—including eSource, EDC, eTMF, and CTMS—the pilot must assess how AI-enabled approaches perform across these varied technology environments rather than relying on isolated or highly standardized data ecosystems alone. Broad participation across these variables will ensure that findings are generalizable and applicable across the full range of clinical trial conditions.

5. Trustworthiness (Aligned with NIST AI RMF)

Our recommendations align with principles reflected in the NIST AI Risk Management Framework (RMF). Evaluation of AI-enabled approaches should occur within an operational context, prioritizing transparency, robust governance, and continuous human oversight to ensure reliability across diverse clinical trial environments.

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

5(a) What evidence should demonstrate that AI systems are valid and reliable in clinical trial contexts?
Evidence of validity and reliability should include a documented validation framework and lifecycle artifacts, demonstrating that the AI system meets predefined performance and risk-based criteria. This includes a validation plan and protocol (with defined metrics and acceptance thresholds), validation reports with test results and sign-offs, and a traceability matrix linking requirements to outcomes, ensuring reproducibility and auditability. Ongoing reliability is supported through independent validation, periodic performance reviews, drift monitoring, and change control documentation, confirming that the system remains fit-for-purpose over time in a regulated clinical context.

Evidence of validity and reliability should include both operational and system-level evaluation within clinical trial workflows. FDA should consider whether AI-enabled approaches demonstrate consistent performance across sites, patient populations, and data conditions representative of clinical trial execution.

Evaluation should also prioritize the quality, completeness, provenance, and traceability of underlying data, along with the ability to document inputs, outputs, workflows, and oversight processes. Demonstrating trustworthiness requires assessing not only the AI-enabled approach itself but also the specific operational systems, workflows, and governance structures in which it is deployed. Maintaining transparency in how these approaches are evaluated and monitored will support confidence among regulators, sponsors, investigators, and research sites, ensuring the integrity of medical evidence resulting from AI-enabled clinical trial workflows.

5(b) How should safety and risk mitigation associated with AI systems be evaluated?
Safety and risk mitigation should be evaluated through a combination of technical, operational, and governance-based controls within clinical trial workflows. Evaluation should prioritize the ability to identify, escalate, review, and resolve errors, inconsistencies, or potential safety issues in a timely and traceable manner.

Layered technical and operational controls – including access management, logging, defined human-oversight requirements, rollback and runbook testing, and periodic control-effectiveness checks – should also be evaluated, aligned with recognized AI risk-management and healthcare cybersecurity standards.

FDA should recognize that risks arise not only from model performance but also from fragmented workflows, inconsistent data quality, interoperability limitations, and incomplete reporting across clinical research systems. Evaluation frameworks should therefore emphasize system-level and workflow-related risks that impact how AI-enabled approaches perform in practice. Furthermore, participants must maintain robust governance processes, oversight mechanisms, change management controls, and ongoing monitoring throughout the pilot lifecycle to ensure the continued safety and integrity of trial operations.

5(c) What metrics can assess transparency and explainability for different stakeholders, and are there metrics available that would be applicable to both sponsor-developed and proprietary systems?
Transparency and explainability can be assessed through a consistent set of cross-system metrics applicable to both sponsor-developed and proprietary AI solutions. These include:

Explainability metrics:

- Degree to which outputs can be traced back to source data, rules, or supporting evidence, and clarity of rationale provided to users

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

Transparency and documentation metrics:

Availability and completeness of model documentation, intended use, limitations, and decision logic, along with audit trails linking inputs to outputs

User understanding and trust indicators:

Stakeholder feedback on interpretability, usability of outputs, and confidence in decision support

Auditability and traceability:

Ability to reproduce results, track data lineage, and maintain logs for regulatory or quality review

Consistency across contexts:

Stability of explanations and outputs across different users, studies, and environments, ensuring reliable interpretation

Transparency and explainability should be evaluated based on whether stakeholders can understand, review, and oversee how AI-enabled outputs are generated and used within clinical trial workflows. Relevant metrics include the quality of documentation regarding intended use, data provenance, model limitations, and oversight processes. Furthermore, auditability and the traceability of operational decisions are essential measures of transparency.

For proprietary systems, independent third-party evaluator attestations against pre-specified performance and documentation standards can supplement public disclosure, preserving IP while supporting regulator and sponsor confidence.

While technical implementation may differ between sponsor-developed and proprietary systems, FDA should maintain consistent expectations for documentation, governance, and traceability across all platforms. Applying these standardized transparency metrics ensures reproducible evaluation and supports broader confidence in AI-enabled implementations, regardless of the underlying technology environment.

5(d) How should privacy protections and data governance practices be evaluated?

Privacy and data governance practices should be evaluated through a comprehensive, risk-based framework that assesses controls across the full data lifecycle—including classification, access, processing, storage, and disposal. This includes verifying data classification and labeling standards, access controls (e.g., least privilege, identity management), encryption and masking mechanisms, and data minimization practices, ensuring data is protected based on sensitivity and regulatory requirements.

Evaluation should also include auditability and monitoring capabilities (e.g., access logs, anomaly detection, and periodic reviews), as well as alignment to regulatory frameworks (e.g., GDPR, HIPAA) and internal governance policies, supported by documented roles (data owners, stewards) and accountability structures.

Privacy protections and data governance practices should focus on validating the transmission mechanics of the pre-agreed JSON reporting schemes, leveraging the pilot's signal-only design to maintain absolute compliance without exposing patient-level data.

Because clinical trial data often originates from fragmented systems, FDA should prioritize how data is aggregated, standardized, and integrated across operational environments. Robust governance practices are

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

foundational; they support not only privacy protection but also the underlying reliability and trustworthiness of AI-enabled approaches across the clinical trial ecosystem.

5(e) What approaches should be used to assess fairness across demographic and clinical subgroups?
The paramount consideration here is that fairness evaluation should include pre-specified subgroup performance metrics and disparity thresholds, with defined mitigation steps – such as workflow safeguards or threshold adjustments – when clinically meaningful differences are observed. It should also be explicitly acknowledged that statistical power for subgroup analyses will be limited in early-phase trials with small populations; the pilot should require their prospective measurement while treating subgroup performance as hypothesis-generating rather than definitive.

Fairness should be assessed through systematic evaluation across diverse demographic groups, patient populations, therapeutic areas, geographic regions, and operational settings representative of the clinical trial ecosystem. Consistent with NIST AI RMF principles, assessment should determine whether AI-enabled approaches demonstrate consistent and reproducible performance across these variables to prevent biased outcomes in trial participation or data interpretation.

Evaluation should utilize stratification across site types (e.g., academic medical centers vs. community-based sites), study phases, and geographic regions to identify potential performance disparities. Ensuring fairness requires looking beyond model behavior to address structural factors, such as data quality, site-level operational maturity, and fragmented workflow variability, which can implicitly encode bias. By prioritizing standardized technology platforms that connect sponsors and sites, the industry can better ensure equitable access to clinical research opportunities and verify that AI-enabled improvements are scalable and fair across all populations.

6. Comparative Evaluation
6(a) What comparators are most appropriate (e.g., historical controls, concurrent non-AI trials, simulation studies)?
Within-trial designs (for example, site-level stepped-wedge rollout or pre/post implementation) are often efficient for operational use cases. In-silico simulation, digital twins, and synthetic data should be treated as complementary evidence to explore counterfactuals and stress-test performance, not as standalone substitutes for patient-level evidence.

We would also suggest a layered comparison hierarchy: for each AI tool, the *primary* comparison should be against the pre-specified performance threshold in the credibility assessment plan, not against an external comparator. There should be a *secondary* comparison against the trial's own historical baseline. There should be a *tertiary* comparison against matched concurrent trials or simulation.

6(b) How should differences in trial design, complexity, or therapeutic area be accounted for in comparisons?
Differences in trial design, complexity, and therapeutic area should be accounted for through risk-adjusted and stratified comparisons, ensuring like-for-like evaluation. This includes normalizing key metrics by study characteristics (e.g., phase, population size, protocol complexity) and conducting subgroup analyses within similar therapeutic areas or trial types. Comparisons should also control for operational variables (e.g., site mix, geography, enrollment dynamics) to isolate the impact of AI.

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

Differences in trial design, complexity, and therapeutic area should be addressed through stratified, context-specific evaluation rather than relying solely on aggregate comparisons or broad normalization. Evaluation frameworks should stratify by site type, study phase, geographic region, and operational maturity to enable meaningful comparisons across similar conditions. Where AI may alter participant selection or endpoint measurement, comparisons should establish a clear counterfactual and align estimands for population, endpoints, and follow-up, consistent with ICH E9(R1) estimand principles.

Because statistical adjustments alone cannot fully account for variability in operational infrastructure, workflow integration, and data interoperability across organizations, findings must be interpreted within the specific context of the execution environment. Broad participation across diverse sponsors, research settings, and technology ecosystems is necessary to ensure that pilot findings are reproducible, generalizable, and operationally meaningful across the full spectrum of clinical trial conditions.

7. Qualitative Outcomes

7(a) How can stakeholder trust in AI-enabled trial approaches be assessed (e.g., investigators, participants, regulators)?

Stakeholder trust in AI-enabled trial approaches should be assessed through a combination of adoption, transparency, and outcome-based metrics, alongside structured feedback. This includes measuring user adoption and engagement (e.g., investigator utilization), satisfaction and confidence surveys, and qualitative feedback from investigators, participants, and regulators. Trust is further validated through transparency and auditability (clear documentation, explainability, and traceability of AI outputs) and by demonstrating consistent, high-quality outcomes (e.g., improved efficiency, maintained safety, reduced errors). Continuous monitoring of feedback, escalation of concerns, and visible governance oversight reinforce sustained trust across all stakeholder groups.

Stakeholder trust should be assessed through a combination of operational, qualitative, and experience-based measures. Crucially, evaluation must recognize that patient trust is the foundational requirement for the continued viability of clinical research. If AI-enabled approaches do not maintain participant confidence, the resulting impact on recruitment and retention could undermine the integrity of the entire research ecosystem.

- Participants: Feedback on data stewardship, transparency regarding how AI influences their trial experience, and whether AI-enabled tools improve accessibility and communication rather than creating digital barriers.
- Investigators and Site Staff: Willingness to adopt and consistently use AI tools within day-to-day operations, ensuring these tools reduce rather than increase operational burden.
- Regulators: Confidence in the auditability, traceability, and reproducibility of AI-enabled outcomes.

Ultimately, trust must be evaluated through demonstrated consistency and reliability. Maintaining a transparent "human-in-the-loop" oversight model ensures that AI serves to enhance, rather than replace, the trusted relationship between participants, investigators, and the clinical research community.

7(b) What methods can evaluate usability and integration into clinical workflows?

Usability and integration into clinical workflows should be evaluated through a combination of user-centered metrics, workflow impact analysis, and evidence-based change management drivers. This includes measuring user adoption, task efficiency, and workflow fit (e.g., time to complete tasks, reduction in manual steps), alongside structured user feedback and satisfaction assessments to validate ease of use and relevance.

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

In addition, evaluation should incorporate behavior change indicators—such as adoption rates, sustained usage, and alignment with defined "critical behaviors"—to confirm that AI is not only used but embedded into routine practice, supported by training, reinforcement, and feedback loops. Finally, process-level metrics (e.g., cycle time reduction, error reduction, and consistency of outputs) demonstrate whether the AI solution meaningfully improves workflow performance while integrating seamlessly into existing clinical operations.

Usability and integration should be evaluated by how effectively AI-enabled approaches operate within existing trial processes and operational environments. Evaluation should prioritize quantifiable reductions in manual effort, cycle times, duplicate documentation, and the reliance on "workarounds" that currently characterize fragmented research settings.

Assessment must also consider how AI integrates with existing technologies, including "real-world" environments where hybrid or paper-based processes remain common. Because the mere presence of AI functionality does not guarantee adoption, successful integration depends on whether these approaches align with day-to-day workflows and reduce administrative "noise." Feedback from investigators, site staff, and study coordinators is essential to determine if AI-enabled tools meaningfully improve trial execution. Ultimately, by reducing operational burden and workflow disruption, these tools allow site staff to dedicate more focus to the patient experience and clinical care.

7(c) How should perceived value, scalability, and operational feasibility be measured?
Perceived value, scalability, and operational feasibility should be measured through a combination of outcome-based, adoption, and capability metrics. This includes assessing value realization (e.g., time saved, cycle time reduction, productivity gains), alongside adoption and utilization rates to confirm that the solution delivers meaningful impact in practice. Scalability is evaluated through the ability to replicate performance across studies, sites, and workflows, supported by consistent results and infrastructure readiness. Operational feasibility is assessed by measuring workflow integration, data readiness, and resource requirements, ensuring the solution can be reliably deployed, governed, and sustained within existing clinical operations.

Perceived value, scalability, and operational feasibility should be measured by whether AI-enabled approaches deliver quantifiable operational improvements across diverse research settings. Relevant measures include reductions in time-to-completion, administrative burden, duplicate processes, and manual review activities, alongside improvements in workflow consistency and interoperability. Practical scalability indicators should also include FTE burden per enrolled participant and cost per activated site, with a clear distinction between nominal tool availability and active workflow use.

Evaluation must reflect the reality of the fragmented clinical trial ecosystem. Approaches that demonstrate value in resource-rich or highly standardized environments may not scale across operationally diverse settings. Therefore, FDA should assess whether AI can scale without requiring excessive customization, increased operational overhead, or total workflow redesign. Scalability depends on technical performance paired with robust governance and the ability to operate consistently across fragmented systems (e.g., eSource, EDC, CTMS). Ultimately, operational feasibility should be judged by whether AI-enabled approaches support a sustainable and reliable modernization of trial execution that remains accessible to all patient populations and research settings.

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

We support FDA's efforts to evaluate AI-enabled approaches in clinical research. However, the success of AI in clinical trials will depend far less on model sophistication alone than on the modernization of the shared, global operational infrastructure supporting trial execution and regulatory data exchange. ACRO members are ecosystem enablers and possess the requisite expertise to support reference evaluation frameworks and pilot cohorts.

Clinical research remains fragmented across disconnected systems, inconsistent workflows, and duplicative processes. These operational limitations—rather than limitations in AI itself—represent the primary barrier to scalable improvements in trial efficiency and quality. The proposed pilot presents an opportunity not only to evaluate AI capabilities but to determine whether more interoperable, structured, and machine-readable approaches can reduce friction across the broader clinical research ecosystem.

FDA should prioritize evaluation methods that reflect the operational realities of research, including varying levels of digital maturity and hybrid operational models. Prioritizing interoperable infrastructure and structured data exchange will deliver greater long-term impact than the isolated automation of individual workflow steps. Ultimately, successful adoption of AI will depend on transparency, governance, and the industry's confidence in the resulting medical evidence. FDA has a unique opportunity to help shape not only how AI is evaluated but how the underlying clinical trial ecosystem evolves to support more scalable and efficient development.

Section Three – Questions That Were Not Posed in the RFI:

We would like to address a handful of questions that were not actually posed in the RFI, but may be useful:

What kind of governance model would best support FDA's real-time signal review?
The architecture contemplated by the pilot gives FDA (near or real time) continuous visibility into trial data, which raises the question: What happens if FDA's analytical systems identify a safety signal before the sponsor or investigator has done so?

Without a clearly defined protocol governing this scenario, sponsors would face genuine ambiguity around accountability and regulatory exposure—which, if unaddressed, could have a chilling effect on pilot participation, particularly among the data-mature sponsors whose involvement is most valuable. We ask FDA to consider publishing a signal response protocol prior to pilot launch, along with a safe harbor provision for sponsors who respond promptly and in good faith following agency notification.

How should the pilot account for and address existing infrastructure barriers and varying digital maturity at the research site level to ensure scalable AI execution?
Realizing AI's full potential requires broad ecosystem modernization. Research sites—including academic medical centers, community-based clinics, and independent sites—operate with highly varying levels of digital maturity, with many still relying on paper-based workflows. To establish a foundation for scalable AI, the FDA and the Department of Health & Human Services should actively encourage the adoption of standards-based technologies like direct data capture and digitally enabled workflows across all sites.

How must regulatory submission frameworks evolve to allow both humans and automated systems to efficiently review clinical trial information?

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org

Clinical trial submission packages have grown so large and complex that they are increasingly difficult for both humans and machines to parse. Frameworks must be modernized by structuring regulatory and clinical trial information through interoperable, machine-readable, and standards-based approaches. This shift will optimize review efficiency, lower the operational burden on stakeholders, and allow for scalable oversight of AI-enabled programs.

Can AI-enabled technology approaches be evaluated centrally at a program or platform level to avoid redundant, trial-by-trial technical reviews?

Yes, the FDA should implement a model where certain operational capabilities or AI methodologies (such as real-time anomaly detection or data quality monitoring) undergo a single, foundational evaluation of governance and methodology. Once validated at this platform level, sponsors could implement the approach across their entire clinical portfolio without needing duplicative de novo technical reviews for every individual protocol, thereby reducing redundancy and increasing consistency. However, even if a platform-level review model is used, sponsors would still need to confirm the AI is appropriate for the specific trial and context where it is being applied.

What specific site-level workflow automations or data-capture standards are required to achieve a 24-hour data transmission window without increasing manual data-entry duplication for clinical research coordinators?

To eliminate expensive, study-specific programming and ensure scalable mapping across disparate source architectures, sponsors and technology partners must utilize certified technology platforms that natively support discrete data standards. Certified platforms establish a centralized data architecture that automatically normalizes incoming source information into discrete, standardized data elements—such as HL7 FHIR schemas—at the point of ingestion. This standardized foundation allows a configurable semantic mapping layer within the technology to dynamically translate these definitions directly into the agency's expected JSON metadata formats. Consequently, sponsors can instantly initialize new trials because the underlying validated software automatically maps, abstracts, and routes operational and safety signals without bespoke software development.

We thank the FDA for this opportunity to provide feedback and look forward to continued collaboration with FDA and industry stakeholders as these efforts evolve.

Respectfully submitted,

Karen Noonan

Karen Noonan, Senior Vice President, Global Regulatory Policy

601 New Jersey Ave NW, Suite 350, Washington DC 20001 | info@acrohealth.org | www.acrohealth.org