Paper ML28

From Demo to Deployment: Bridging the AI Implementation Gap

Melanie Hullings, Formation Bio, New York, NY, US
Emily Yates, Formation Bio, New York, NY, US

ABSTRACT

The pharmaceutical industry is at a point in time where many are trying to translate impressive open-source AI demonstrations into production-ready tools that deliver measurable impact. In a regulated environment, some skepticism is justified. While flashy demos showcase potential for drug discovery and clinical optimization, practical implementation in a GxP setting requires systematic approaches prioritizing reliability, interpretability, and compliance.

Successful transformation demands rigorous validation frameworks that evaluate real-world performance under diverse conditions, robust data governance protocols, and continuous monitoring systems. Responsible AI adoption requires cross-functional collaboration between data scientists, domain experts, and regulatory teams from early development stages, embedding strict system requirements based on intended use and ethical principles - fairness, accountability, and patient safety - into core design processes.

In this talk, you will learn practical frameworks for evaluating AI readiness, strategies for building cross-functional implementation teams, and methodologies for ensuring quality output while maintaining innovation velocity in pharmaceutical AI deployments.

INTRODUCTION

Artificial intelligence has become increasingly visible across pharmaceutical research and development, with widespread interest in generative AI, large language models (LLMs), and agentic systems. Industry surveys indicate that while AI adoption is broad, most implementations remain in early or pilot phases and are not yet embedded into mission-critical, regulated workflows (McKinsey & Company, 2025). As a result, despite compelling demonstrations, core clinical trial timelines such as study startup, database lock, and submission preparation have not substantially improved.

In regulated environments, skepticism toward AI is warranted. Clinical development processes are highly interdependent, quality-sensitive, and subject to regulatory scrutiny. AI systems that lack interpretability, reproducibility, or clearly defined intended use can introduce risk rather than reduce it. Consequently, successful AI implementation in GxP settings requires more than model performance; it demands careful alignment with process design, decision accountability, and risk-based quality principles.

This paper addresses a fundamental implementation question: **how should organizations identify where AI can have meaningful impact on a clinical trial, and how should such systems be designed to deliver measurable improvements without compromising quality?** Rather than focusing on AI as a technology, we focus on AI as a decision-support mechanism that, when properly designed, can move high-risk decisions earlier in the clinical lifecycle.

AI READINESS AND THE LIMITATIONS OF SECONDARY WORKFLOWS

Most current AI applications in clinical development function to support or augment existing **individual or exploratory workflows** without altering when primary decisions occur. Examples include AI-generated protocol summaries, automated QC suggestions, or draft document generation. These applications can improve efficiency or reduce manual effort, but failure of the AI does not typically delay the trial, as the primary workflow proceeds unchanged.

While valuable, **secondary workflows** rarely shift critical path timelines. Industry analyses have shown that only a minority of AI initiatives scale beyond experimentation into production systems with measurable operational impact [1]. This gap between intent and impact highlights the need for AI readiness frameworks that distinguish between productivity gains and true process transformation.

In contrast, **primary or critical-path AI workflows** directly influence when downstream activities can begin or complete. In regulated environments, these workflows must be designed with explicit quality gates, interpretability, and fallback mechanisms to ensure trust and compliance.

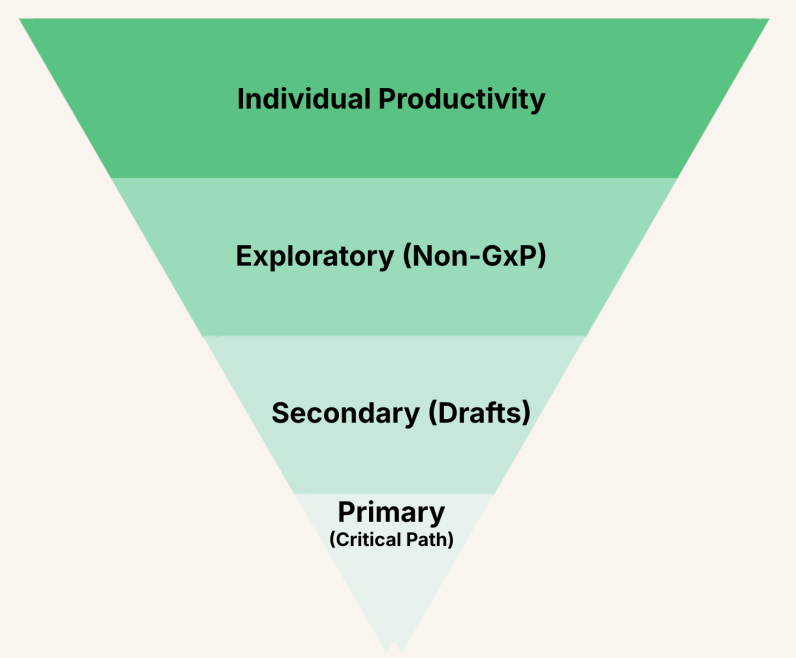

Figure 1: Current evaluation of AI readiness

IDENTIFYING A HIGH-IMPACT USE CASE: EDC BUILD IN STUDY STARTUP

To identify a candidate use case for high-impact AI implementation, we began with first principles evaluation of critical path timelines and impactful processes. Study startup was selected as an initial focus due to its frequent role as a bottleneck and its strong influence on overall development timelines. The meaningful bottleneck where AI implementation can impact these timelines then became the area of focus for critical path implementation.

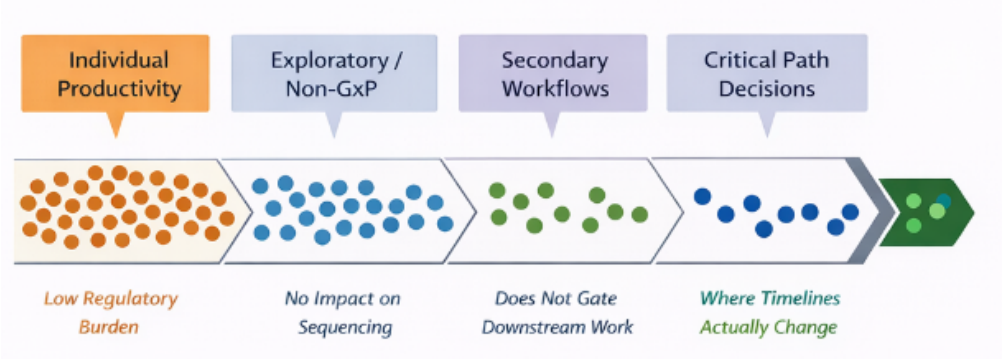

Figure 2: Focus on impact over individual workflows to implement AI on the critical path

Within study startup, EDC build emerged as a recurring constraint. Industry-standard EDC builds often require approximately 12 weeks and rely heavily on manual interpretation of protocol documents, multiple rounds of cross-functional review, and late discovery of inconsistencies. While industry automation efforts have historically emphasized standards-based approaches, standardization alone cannot capture protocol-specific clinical nuance. Loss of nuance often results in extensive CRF redesign during study startup, mid-study CRF updates, amendments, and downstream rework, increasing cost and delaying execution.

Based on these observations, the guiding hypothesis was: **if AI could reduce late-stage rework by enabling earlier, higher-quality interpretation of protocol requirements, EDC build timelines could be materially shortened without sacrificing quality.**

INITIAL APPROACH AND FAILURE OF ONE-SHOT AUTOMATION

The initial implementation attempted to automate EDC specification generation end-to-end. Given a protocol and schedule of assessments, the AI system produced a monolithic EDC specification intended for direct use by EDC builders or upload into an EDC platform.

Although technically impressive, this approach failed to deliver operational confidence. The generated specification was difficult for clinical data managers to review, lacked sufficient transparency into decision logic, and provided no clear mechanism for assessing whether the output was “good enough” to support aggressive timeline commitments. Additionally, attempts to mirror traditional user acceptance testing (UAT) by incorporating multiple functional reviewers introduced overlapping and sometimes conflicting feedback, increasing system complexity without improving clarity.

From a system-design perspective, this approach effectively created an uncoordinated multi-objective optimization problem. The AI was implicitly asked to satisfy numerous reviewer perspectives simultaneously without a clear prioritization or arbitration framework. This experience underscored a critical insight: **accuracy alone is insufficient in GxP AI systems; interpretability and decision transparency are prerequisites for trust.**

DESIGN PIVOT: COMPONENTIZATION AND RISK-WEIGHTED REVIEW

The successful redesign began by decomposing the EDC specification into discrete, reviewable components, including visit schedules, forms, protocol citations, fields, and codelists. Each component was evaluated based on its downstream impact on rework risk. Errors in assessment and form mapping, for example, were found to invalidate large portions of downstream work, whereas errors in lower-level elements such as field formatting were less costly to correct later.

By organizing generation and review around components rather than complete artifacts, the workflow enabled sequential, risk-weighted review. High-impact decisions were reviewed and resolved before downstream components were generated, reducing the propagation of errors. Importantly, this approach aligned AI output with human cognitive workflows, allowing domain experts to apply judgment where it mattered most.

Protocol citations were introduced as a critical interpretability feature, enabling reviewers to evaluate AI-generated components directly in the context of relevant protocol language. This transformed the system from an opaque generator into a decision-support tool that accelerated expert reasoning rather than replacing it.

MOVING WORK UPSTREAM: FROM UAT TO PROTOCOL DEVELOPMENT

An unexpected but significant outcome of this redesign was the realization that the EDC Spec Builder could be used earlier than originally intended. By enabling structured interrogation of protocol requirements through the lens of data collection, many discussions that traditionally occurred during UAT were shifted upstream into protocol development.

This shift reframed the role of EDC build from a downstream execution task to an early planning and quality-assurance mechanism. In practice, this resulted in substantial protocol clarifications, clinical and operational recommendations, and resolution of ambiguities prior to regulatory submission. Early use on an active protocol demonstrated tangible improvements in protocol clarity and readiness, with potential to reduce amendments and mid-study changes.

This outcome reinforced a broader principle: **AI delivers the greatest value when it enables teams to think better and faster about high-risk decisions, not when it attempts to replace expert judgment.**

RISK-BASED TESTING AND TRUST IN GXP AI SYSTEMS

Traditional software validation approaches based on deterministic requirements and test cases are insufficient for AI systems. Non-deterministic outputs, multiple acceptable solutions, and evolving models necessitate a risk-based evaluation framework aligned with ICH E6(R3) data integrity and ICH Q9 quality risk management principles, as well as emerging AI guidance (ICH, 2023; FDA, 2026).

For the EDC Spec Builder, quality was assessed using component-level metrics, severity-weighted error classification, and iterative gap-filling review loops. Rather than optimizing for model accuracy in isolation, testing focused on outcomes that directly influenced timelines, such as prevention of rework and early detection of protocol issues. Release hygiene, versioning, and documented decision heuristics further supported operational trust.

A TRANSFERABLE FRAMEWORK FOR AI IMPLEMENTATION IN CLINICAL DEVELOPMENT

Based on this experience, a practical framework for AI readiness in GxP environments emerges:

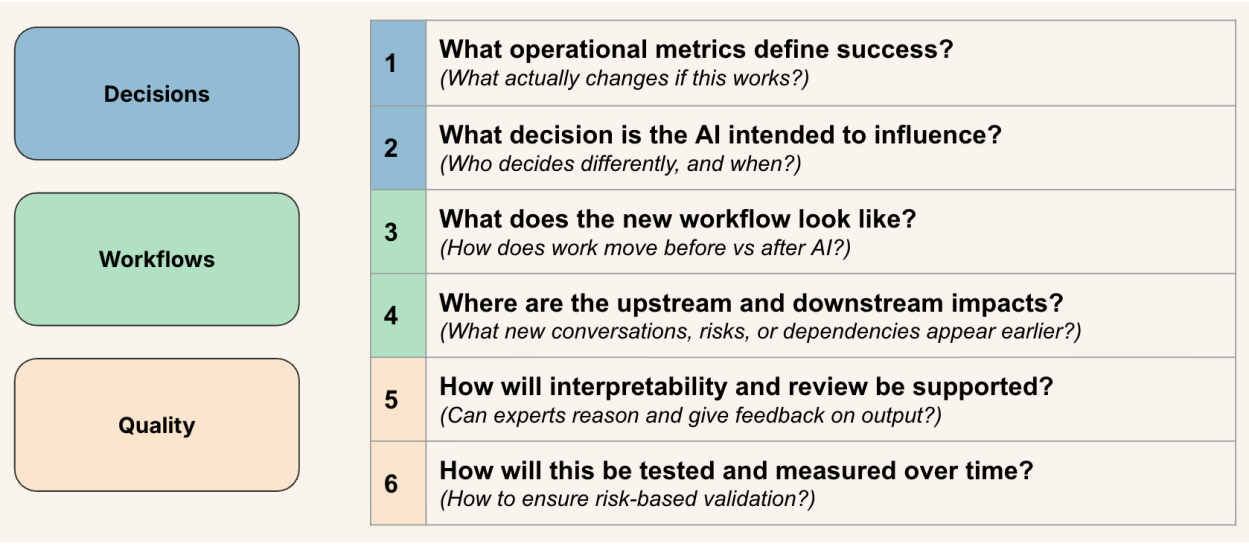

Figure 3: Practical AI implementation framework to realize impact

This framework is applicable beyond EDC build for clinical data management, including protocol design review, edit check logic evaluation, vendor data integration, and database lock readiness assessment. Programming and other heavy process and review-based use cases in the GxP are also opportunities to continue to explore further within this framework to unlock new workflows.

CONCLUSION

Recent regulatory guidance emphasizes that successful application of artificial intelligence in drug development depends not only on model performance, but on clearly defined intended use, human oversight, transparency, and continuous lifecycle management. The U.S. Food and Drug Administration's guiding principles for good AI practice underscore the importance of interpretability, risk-based validation, and ongoing monitoring when AI systems are used to support decision-making in regulated environments (FDA, 2026).

The framework presented in this paper operationalizes these principles at the workflow level. By beginning with measurable operational outcomes, explicitly identifying the decisions being influenced, redesigning workflows to surface risk earlier, and embedding interpretability and review into system design, the framework aligns closely with FDA expectations while remaining practical for real-world clinical development. Importantly, testing and measurement are treated as continuous activities rather than one-time validation exercises, reflecting the lifecycle approach described in regulatory guidance.

While the case study described focuses on EDC build and protocol development, the framework is intentionally transferable across other GxP workflows, including programming, quality control, vendor data integration, and database lock readiness. In each case, the central lesson remains consistent: artificial intelligence delivers meaningful value when it accelerates expert thinking and moves high-risk decisions earlier in the clinical lifecycle, without diminishing accountability or transparency.

As organizations continue to explore AI adoption, aligning implementation strategies with both regulatory guidance and operational realities will be critical. Frameworks that emphasize decision timing, workflow redesign, and risk-based oversight offer a path toward realizing the potential of AI while maintaining the rigor required in regulated clinical development.

REFERENCES

1. Singla, A., Sukharevsky, A., Yee, L., Chui, M., Hall, B., & Balakrishnan, T. (2025). The state of AI: How organizations are rewiring to capture value (McKinsey Global Survey on AI). McKinsey & Company. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai
2. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. (2023). *ICH guideline E6(R3): Good clinical practice*. https://www.ich.org/page/efficacy-guidelines
3. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. (2023). *ICH guideline Q9(R1): Quality risk management*. https://www.ich.org/page/quality-guidelines
4. U.S. Food and Drug Administration. (2026, January). Guiding principles of good AI practice in drug development (FDA Doc. No. 189581). https://www.fda.gov/media/189581/download

ACKNOWLEDGMENTS

We would like to acknowledge the cross-functional Formation Bio team whose collaboration across data management, engineering, product, and clinical development made this work possible.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Author Name: Melanie Hullings
Company: Formation Bio
Email: melanie@formation.bio

Author Name: Emily Yates
Company: Formation Bio
Email: eyates@formation.bio