The Drug Got a Supercomputer.
The Trial Got a Spreadsheet.

Ashley Essick, MBA, PMP | Global Project Manager, ICON PLC

The pharmaceutical industry has made an extraordinary bet on compute. Eli Lilly and NVIDIA announced a billion-dollar co-innovation lab in January 2026. Bristol-Myers Squibb deployed NVIDIA's DGX SuperPOD for oncology R&D and reported 55% infrastructure cost savings. AstraZeneca helped build Cambridge-1. Amgen placed a supercomputer in Iceland. The industry has signaled that the fundamental approach to drug development has changed and how we are discovering new molecules is seeing the greatest impact currently. AI is compressing early discovery timelines by 30 to 40 percent. Preclinical candidate development that once took three to four years is reaching readiness in 13 to 18 months.

The progress we are seeing is phenomenal, but the molecule still has to survive the trial.

The clinical execution layer is running on infrastructure that has not kept pace. The same executive teams authorizing supercomputers to screen billions of compounds are still

managing Phase III execution through retrospective monitoring cycles, manual query workflows, and periodic data reviews. As an industry, we are utilizing the most advanced compute power in history to determine which drug molecules are most likely to prove effective, while the operational models used to demonstrate efficacy and endpoints have not yet evolved at the same pace. The industry is beginning to recognize this gap and the organizations moving first will define the standard. The discordance in these models presents us with an opportunity to innovate and unleash the full power of agentic AI to improve the total drug development process from discovery to approval.

That gap is a technological opportunity where capital allocation can equally demonstrate both cost and time wins at a greater rate than previously calculated.

The Economics of Execution

Developing a single approved drug costs between $708 million and $2.8 billion, with the median closer to $1.3 billion when you account for cost of capital and failed candidates. Phase III is the most expensive single stage, averaging $350 million and reaching $1 billion for complex programs. At that scale, operational delays are not inconveniences. They are substantial financial events which not only impact drug cost but ultimately delay critical therapies from having impact on the lives of patients and families.

Every day of Phase III delay costs approximately $500,000 in unrealized commercial value plus $40,000 in direct operational costs. Delays in Phase III are not rare exceptions. Rather, they are structural features we accept, and sometimes even budget for, with limited innovative thinking about how we can proactively anticipate those occurrences most likely to cause delay.

The Tufts Center for the Study of Drug Development has documented this clearly. Seventy-six percent of Phase III protocols now require at least one substantial amendment, with a mean of 3.5 amendments per Phase III protocol. Each amendment takes an average of 260 days to implement from the moment the need is identified to the last oversight approval. During that window, sites operate under different versions of the protocol for 215 of those days. The direct cost of implementing a single substantial Phase III amendment is $535,000. Total indirect costs run three to four times higher.

Meanwhile, TransCelerate BioPharma and Tufts CSDD documented something equally striking: Phase III pivotal trials have seen a 40% increase in total procedures and a 283%

increase in data points collected over the past decade. Phase III protocols now average 5.96 million data points. Nearly one-third of those data points come from procedures that do not support primary or key secondary endpoints. The execution environment has become more data-dense, more fragmented, and more operationally complex precisely as the tools for managing it have remained fundamentally unchanged.

This is the environment where a Phase III team is expected to catch protocol deviation trends, surface diversity gaps, monitor site performance across dozens of sites spanning multiple continents, manage query rates on data collected years ago, and ensure every evaluable patient reaches every required timepoint. All of this is typically managed through periodic reports and manual tracking workflows.

It does not work. The data confirms it through late-stage rework, incremental cost in time and operational resources, and delayed approvals.

Already Built. Not Yet Deployed.

There is a positive outlook here because the compute architecture to address clinical trial operational challenges already exists and is deployed at scale in adjacent domains. Our priority now must be to invest and implement this layer at scale.

NVIDIA AI Foundry is the intelligence-building layer and the foundation of what is operationally possible in clinical trial execution. It trains domain-specific AI models on proprietary clinical and operational data: query rates, site performance variance, enrollment patterns, protocol deviation clustering, sample logistics requirements. Think of this as the place where historical trial experience gets transformed into predictive models tuned to the specific language and logic of clinical execution.

NVIDIA NIM microservices are then the execution layer. Once those models are trained, NIM deploys them as fast, containerized inference services accessible to operational workflows via real-time API inference, not batch cycles. The result is an intelligence layer that observes the trial continuously and surfaces signals as they emerge, rather than after the review cycle closes. Furthermore, real-time review results in consistency of governance and management of trials, delivering a higher level of data quality and TMF integrity throughout the study.

Together, these two components represent a shift in what clinical operations can see and when. This is more than a dashboard upgrade. It is an architectural change in how a trial generates and responds to operational intelligence. It takes trial management from slow and deviation-prone reactive models to proactive intelligent operations that deliver quality data and endpoints which withstand regulatory scrutiny.

This infrastructure is already demonstrating its potential in adjacent clinical domains. ConcertAI, operating the largest curated oncology data repository in the industry at more than 8 million patients, integrated NVIDIA NIM microservices to support clinical trial patient matching, protocol automation, and research site co-pilots with real-time analytics. The University of Florida Health and NVIDIA built GatorTron, a clinical NLP model trained on 10 years of data from over 2 million patients, built on the recognition that 80% of valuable clinical information is trapped in unstructured records. The result: identifying candidates for clinical trials went from months to minutes.
At NVIDIA's GTC 2026 conference, Jensen Huang described AI agents deployed across the clinical research ecosystem as compressing data review cycles from seven weeks to as little as two weeks, with agentic platforms now actively serving 19 of the top 20 pharmaceutical companies.

While these specific operational applications are emerging rather than fully deployed at scale, the underlying infrastructure has been validated in adjacent clinical domains which lends clear support to the architectural path. It is time for the industry to apply this infrastructure to the clinical execution layer where the financial stakes are highest.

Where Compute Meets the Trial Floor

I have spent 13 years managing global clinical trials across Phase I through III, across oncology, rare disease, endocrine and global programs spanning more than 23 countries. I have directly contributed operationally to three FDA-approved therapies. There are patterns in where clinical trial operational delivery breaks down. These repeat across programs, sponsors, and therapeutic areas, and they share a common root cause: data that exists but arrives too late to act on, requiring reactive response where proactive intelligence could have prevented the situation entirely.

The Retrospective Data Crisis

On a global Phase III program, the team received an FDA Information Request mid-review requiring additional patient response data at specific timepoints not originally scoped for submission-readiness. What followed was months of unplanned work: tracking patient visits within narrow windows across a large number of countries,

managing sample stability requirements, resolving queries on data locked up to three years earlier, securing investigator sign-off during holiday periods, and managing protocol deviation cascades where related deviations risked reclassification as major. The unplanned extension, estimated at 6 months after data collection and continued review, at $500,000 per day represents approximately $90 million in unrealized commercial value, before direct remediation costs or the burden placed on sites that had already completed their work.

Could the FDA's data need have been anticipated? The data to model that question existed throughout the trial. The infrastructure to read it proactively did not. A continuous intelligence layer, trained on historical regulatory request patterns and trial-specific endpoint exposures, can surface that risk signal before a submission lands on an agency reviewer's desk. This is not prediction in the speculative sense. This is pattern recognition at a scale and speed that periodic manual review cannot replicate.

The Diversity Planning Failure

The FDA's Diversity Action Plans, mandated under FDORA for Phase III programs and required no later than Phase 3 protocol submission, exist because sponsors consistently treat diversity as a late-stage logistics problem rather than a study design decision. In practice, diversity assessment and outreach commonly begins at 70 to 80 percent enrollment complete. By that point, site selection is locked, enrollment windows are closing, and the populations required to be evaluable for regulatory submission are structurally inaccessible. The remediation adds months and millions to programs that could have planned for this at protocol design.

Population modeling and site selection intelligence, run at protocol design stage using AI trained on historical enrollment data, demographic registries, and site performance records, can identify where the required patient populations exist and which sites can actually access them before a single site is selected. This is a planning problem masquerading as a recruitment problem. Compute solves it at the front end, where the cost is informational rather than operational.

The Protocol Design Blind Spot

Protocol structures that create unintended behavioral incentives at the site level often go undetected until the pattern has already compromised data integrity across the enrollment

population, potentially leading the trial itself to fail. Individual site decisions that each appear to be within protocol discretion become a systemic data quality problem when viewed across the full study. By the time this surfaces through manual review in risk or oversight meetings, the damage is substantially done.

Real-time pattern detection across enrollment, randomization, and early withdrawal data, running continuously across all sites, is the only way to surface anomalous behavioral trends before they become a regulatory problem. The current monitoring infrastructure, operating in cycles and reviewing data retrospectively, structurally cannot provide this.

From Retrospective Auditing to Predictive Engineering

Clinical operations have historically been designed to find mistakes after they happen, document them, and remediate. That model was appropriate when data volumes were manageable and regulatory expectations were less exacting. Neither condition applies to modern Phase III execution. This is an unsustainable approach that leads to higher cost across the development process, longer approval timeframes with potential rework, and missed impact for therapies critical to patients.
The TransCelerate and Tufts CSDD research demonstrates the challenge through data: Phase III protocols average 5.96 million data points across increasingly fragmented execution environments spanning global sites, decentralized components, external vendors, and multi-modal data streams. Continuous, real-time operational intelligence is not a luxury in this environment. It is the only rational response to the complexity the industry has already accepted.

The regulatory environment is moving in the same direction. ICH E6(R3) explicitly embraces risk-based quality management and quality by design, which are not achievable with periodic retrospective monitoring. The FDA and EMA jointly released ten guiding principles for good AI practice in drug development in January 2026, establishing a shared framework for AI across the full medicines lifecycle. The FDA's January 2026 Bayesian methodology guidance, endorsing adaptive trial designs that update statistical inferences continuously as trial data accumulates, depends on exactly the compute infrastructure this article describes. The regulatory environment is already shifting how we should be executing and is opening the pathway for intelligent, proactive trial management. The execution infrastructure is now primed for the technological innovation that the complexity of Phase III trials demands.

The Full Pipeline Vision: From Development to Approval

The industry has invested heavily in compute at the discovery layer and the results are compelling. AI is compressing discovery timelines, improving candidate quality, and reducing the attrition that has historically made drug development one of the most capital-inefficient enterprises in the economy. Yet a better molecule still has to navigate a process where clinical phases alone average nearly eight years. A Tufts CSDD and Drug Information Association assessment of 302 drug development professionals across 79 companies found that AI-enabled activities already yield an average 18% cycle time reduction across clinical operations workflows. The infrastructure to extend that impact across the execution layer is available now.

When we model the combined effect of AI applied across the full pipeline, the potential is meaningful: discovery compression of 12 to 18 months, elimination of 3 to 6 months of avoidable operational delay across clinical phases, and improved candidate quality reducing attrition from the industry's 90% failure rate. At $500,000 per day in unrealized commercial value, recovering up to two years at the high end of this range represents approximately $350 million per program - a figure conservatively modeled from published delay benchmarks: 730 days at $500,000 per day equals approximately $365 million, rounded to reflect that not all delay is eliminable and that programs vary significantly in commercial profile. That figure does not count the trials that succeed rather than fail because the signal was caught in time.

No financial model captures what it means to a patient with a progressive disease when a therapy reaches them two years earlier. We cannot fully quantify this impact, and qualitative rhetoric does not touch the true breadth of the effect on patients, families, and communities. It is our responsibility as an industry to implement technological innovations responsibly and efficiently in order to deliver on what we state is a core value: improving the lives of the patients we serve.

The Foundation Is Available Now

The industry has already accepted that compute is essential in discovery. The same logic applies to the execution layer and we have the data to demonstrate its effectiveness. The patterns of operational inefficiency are both known and detectable. The infrastructure to read them in real time exists, has been validated, and is producing results in adjacent domains. What has not yet happened is the deliberate decision to apply it where the money, time, and patient impact is actually being lost.
We can continue treating clinical execution as an administrative function: processing data retrospectively, catching problems after they compound, and absorbing delay costs as a fixed feature of drug development. Or we can recognize clinical trial execution for

what it is: the most data-intensive, highest-stakes operational engineering challenge in life sciences, running on infrastructure that was not designed for its current complexity.

The compute layer to address this is not theoretical. It is not five years away. I encourage innovators and strategic solutions teams across the industry to consider the technological implications of NVIDIA AI Foundry, which builds the domain intelligence, and NIM microservices, which deploy that intelligence where work happens. This is one of our greatest opportunities to improve the functionality and delivery of approved therapies to the market today. The question is simply whether the industry will allocate it where it matters most.

We have the opportunity to start building that foundation now for fully intelligent drug development from discovery to approval.

Examples described are anonymized composite patterns drawn from global clinical trial experience and are not attributable to any specific sponsor or program.

FIRST COMMENT — paste as first comment on LinkedIn post, not in post body:

#ClinicalOperations #AIGovernance #ClinicalTrials #NVIDIA #DrugDevelopment #ClinicalResearch #LifeSciences