What AI Companies Need to Know About Clinical Development

Aaron Smith
Founder, Head of AI
unlearn.ai

Most AI investment in biopharma has gone into drug discovery: target identification, molecule prediction, and drug-screening optimization, for example. This is well-funded territory, and increasingly well-served. Once a promising treatment is discovered, though, there can still be a decade or more of in-human clinical trials, where the treatment faces its ultimate test: proving its efficacy and safety in diverse populations. This is the longer, more expensive part of the process, and where many high-stakes programs fail. AI has played a relatively minor role here so far. That's where useful AI work is.

The obvious AI play is productivity at the SaaS layer: scribing, document drafting, and information flow. These are legitimate ways to speed up trial execution, but shaving years off the process requires tackling the questions clinical trials are designed to answer: *is this treatment effective, is it safe, who does it work for?* That's the evidence regulators and the medical community use to decide whether a treatment moves forward. These are the questions Unlearn works on, alongside the sponsors running those trials.

Domain expertise is structural

Doing AI work in clinical development means closely partnering with clinical development leads, including Heads of Development/Chief Medical Officers, biostatisticians, and regulators. All of them are deep experts in their therapeutic indication and need help from data and models to make complex decisions. Any AI tool brought into their work has to earn their trust and find a clear role in how they make those decisions.

Building tools that earn trust requires an understanding of the disease area. Domain knowledge is the context for every methodological choice. When data must be merged across sources, or scanned for anomalies, every choice must be defensible in the language of the disease. When data is scarce (and it usually is), even basic modeling assumptions are reflections of the domain. Your output will be read by domain experts with years of expectations behind them. If it surprises them, you'd better have a watertight justification in their language.

AI scribe companies employ doctors; Legal AI companies hire lawyers. The same shape applies here. The reflex may be to assume domain knowledge is a fine-tuning problem: train on a few million medical journal articles and you're done. Fine-tuning helps. It won't pick the appropriate statistical method for the trial, curate data in a way that withstands expert scrutiny, or make modeling assumptions that hold up when data are thin. That work is structural and is built into the system around the model, not in the model itself.

The hard part is the data

You'd hope to apply "big data" methods here, but clinical trial data isn't that big when filtered to what's used to generate scientific evidence. After selecting relevant patients, outcomes, and timeframe, you're looking at thousands of records, not billions. To make matters worse, the data is often partitioned across CROs, sponsors, EDCs, central labs, and imaging vendors. Each data originator jealously guards their fiefdom for legitimate reasons of privacy and competitive position. Real improvement in clinical development requires working across data silos.

But the hardest part is data diversity. Every trial, every site, every therapeutic indication can have different interfaces, schemas, missingness conventions, and visit windows. Curation, the work of linking irregular data into a single usable artifact, is the foundational layer that enables analyses such as treatment-effect estimation, safety assessment, and trial simulation. AI-enabled software with frontier reasoning models is now genuinely good at curation, and makes the work scalable.

And scarcity makes modeling assumptions consequential. The naive response to missing data is simple imputation: fill in with the mean or the last observation and proceed. Such choices obscure assumptions and uncertainty. For example, if your outcomes were measured in young women with a specific genetic marker, saying anything about older men without it requires a domain-grounded argument for why the assumption holds. In practice, this means keeping track of modeling assumptions and making uncertainty visible throughout the analysis.

Our Digital Twin Generators – generative models trained on historical patient-level data to produce comprehensive likelihoods of disease course for each patient– exemplify this philosophy. We carefully keep track of the modeling assumptions that are wrapped into the training procedure, and evaluate those assumptions in validation. Likewise, these models quantify uncertainty and make it available to downstream analysis. The result is a defensible treatment of data gaps, and an honest accounting for uncertainty.

The stats layer you can't skip

Clinical data is small relative to internet-scale ML. That makes uncertainty quantification non-negotiable, both ethically and procedurally. No regulator or clinical team is going to accept "we asked the LLM what the treatment effect is."

The right way to bring AI model predictions to critical clinical development decisions is to situate them inside a causal inference framework that handles potential bias and uncertainty. PROCOVA, pioneered by Unlearn, is one such method. It uses digital twins to produce a covariate that adjusts treatment-effect estimates in randomized trials. Doubly-robust ML is another, applied to estimate efficacy in single-arm trials; it combines two modeling approaches in a way that holds up even if one of them is wrong.

These causal inference methods sit on top of the system. Frontier AI sits underneath them, doing what frontier models do well — curation, software development, and stitching irregular environments together. Between the two, smaller-scale generative models (the ones that produce the digital twins) are called by the workflow under controlled conditions. The methods and tools used in the agentic system need to survive both the moment the trial's data is locked for analysis and the years of regulatory review that follow.

The regulatory landscape supports this hierarchy. The FDA's guidance on AI for regulatory decision-making requires that AI applications be employed within operational and statistical frameworks that adequately and defensibly control patient risk. This consideration shapes how we structure the evidence we deliver in all cases.

The agentic workflow is the product

Here's what we've found works:

- Frontier AI at the substrate, powering data curation, software development, and knowledge navigation. This is the work where irregular environments and natural-language tasks reward strong general models.

- Validated methods on top: digital twins, prognostic models, and statistical frameworks for using these models to inform critical decisions in clinical development. These get rolled into reporting tools, with every piece locked, characterized, and defensible.

- A workflow that ties them together: stored domain artifacts, the AI queries, evaluation criteria, human checkpoints, and a connected knowledge graph that explains how everything relates. Together, these make the system defensible and interpretable.

The agentic workflow systemizes and improves on what would otherwise be traditional consulting work — producing the kind of scientific evidence our customers trust, and making Unlearn a tech-enabled services business that scales while retaining the rigor the field expects.

Closing the gap for patients

AI can accelerate life-changing treatments to patients by equipping clinical development teams with the right tools — tools designed for how trials are actually run, fluent in specific disease indications, rigorous by the field's statistical standards, and built for the data that exists. That requires frontier AI at the substrate, validated methods on top, and a scalable workflow that ties the two together with the discipline regulators and clinical teams expect.

Unlearn has been honing this stack and embedding it in clinical development for years. That work includes more than 12 regulatory interactions alongside sponsors with the FDA and EMA, joint scientific work presented with sponsors such as AbbVie and ProJenX, and digital twin models validated across more than 15 disease indications. Our approach is in use today across active sponsor programs in neuroscience, oncology, immunology, and rare disease. The methodology for using digital twins in clinical trials is advancing rapidly, with single-arm and open-label studies leading the way.

The clinical trials that fail are not just lost programs; they are years subtracted from the lives of patients waiting for what those trials might have proven. The work we are describing is about putting those years back.