Clinical Trials Are the New Bottleneck AI Drug Discovery Has Created an Evidence Infrastructure Crisis

APR 01, 2026 • PAID

5 1 Share

Abstract

The core argument: AI has dramatically compressed preclinical drug discovery, l clinical development timelines remain stuck. The bottleneck has shifted from molecule identification to evidence generation. The next durable health-tech companies won't discover drugs. They'll prove they work.

Key claims:

- AI-driven structure-based and generative methods have increased preclinical throughput substantially, pushing more candidates into already-strained develo pipelines

- Long-run clinical success rates only recently began recovering after decades of decline, per 2025 Nature Communications data – meaning the industry hasn't sc translational efficiency at scale

- FDA's 2025 draft guidance on externally controlled trials effectively outlines a technical stack (phenotype normalization, covariate harmonization, temporal alignment, endpoint ontology mapping) that nobody has fully built yet

- A 2025 Nature Medicine TrialTranslator study showed real-world oncology sur is often ~6 months worse than RCT results, and ~1 in 5 real-world patients don't qualify for phase 3 trials

- 2025 FedECA paper in Nature Communications introduces federated external control arms for distributed settings – a direct blueprint for privacy-preserving comparator networks

- TrialGPT (Nature Communications, 2024) and successor systems suggest patie trial matching is solvable, but recruitment alone is too narrow a moat without a broader trial-state architecture

- Five infrastructure layers where durable category-defining companies will like form: comparator infrastructure, phenotype infrastructure, continuous measure infrastructure, model assurance infrastructure, and adaptive protocol infrastruc

Table of Contents

The paradox nobody talks about out lo

Here's a thing that should be more disorienting than it is: the AI-in-biopharma narrative has largely convinced the industry, investors, and the trade press that t hard part of drug development is finding good molecules. Structure-based desig generative chemistry, AlphaFold derivatives, target ID from genomic embedding of it has gotten genuinely impressive. Not vaporware impressive. Actually impre Some of these tools are running live in discovery programs at major sponsors rig now.

The problem is that making the front end of discovery faster is a little like wider the on-ramp to a highway that's already gridlocked. You don't actually get more to their destination any faster. You just fill up the backup.

Clinical development is the backup. It takes, on average, somewhere between six ten years from first-in-human to approval, and the bottleneck during most of that period is not computational. It's not even primarily biological in the narrow sense evidentiary. The industry is slow at assembling the kind of regulatory-grade, cau defensible, generalizable evidence packages that the FDA and its international equivalents actually need to say yes with confidence. And as AI accelerates the fr end, that evidentiary bottleneck becomes more acute, not less.

This is the paradox that people in clinical development talk about quietly but th rarely makes it into the funding narratives for health-tech companies. Everyone to fund the "AI for drug discovery" story because it maps to a familiar venture pattern: scientific insight, clever model, platform play, licensing revenue or acquisition. The evidence generation story is harder to pitch because the produc less romantic. It involves phenotype normalization, external control cohort asser federated data governance, digital twin validation frameworks, and adaptive mas protocol software. None of that fits on a TED slide.

> But that's exactly where the alpha is.

Why this bottleneck exists now

To understand why this is happening now specifically, it helps to separate two different timelines that are running at very different speeds.

The first timeline is discovery throughput. Over the past five or six years, the combination of structure prediction, generative molecular design, and large-scal biological embedding has genuinely changed the rate at which credible drug candidates can be identified. The number of AI-discovered compounds entering clinical trials – while still small relative to the total industry pipeline – is growir

More importantly, the capital and talent flowing into “AI-native” biotech is enor
which means the pipeline of candidates heading toward IND filings is expanding

The second timeline is development infrastructure. This one has barely moved. T
FDA’s average review timeline hasn’t compressed dramatically. Phase 2 and phas
success rates, while having shown some recent improvement after decades of dec
per 2025 Nature Communications analysis, are still deeply sobering – roughly 10
percent of candidates that enter phase 1 ultimately reach approval depending on
therapeutic area. Enrollment velocity is still plagued by the same problems it wa
twenty years ago: sites are overextended, patients are hard to identify, eligibility
criteria are often written for clean populations that don’t exist in the wild, and
sponsors regularly discover late in development that their trial population doesn
look much like the patients who will actually use the drug if it’s approved.

The structural cause of this divergence is that discovery innovation has been dri
computational methods that are relatively cheap to deploy and iterate, while
development innovation requires touching the actual trial infrastructure: sites, I
patient populations, regulatory submissions, comparator data sets, endpoint
definitions. That stuff is slow, bureaucratic, and deeply institutional. It’s not the
of thing you can iterate on quickly with gradient descent.

What makes this moment distinct is that multiple academic and regulatory threa
are now converging on a shared diagnosis, and regulators themselves are beginni
acknowledge that the evidence generation stack needs to be rebuilt rather than
patched. That combination – a growing pipeline of AI-discovered candidates, a
strained development infrastructure, and a regulatory environment signaling
conditional openness to new methodologies – is the setup for a very large busine
opportunity.

The technical stack academia is quietly building

The most interesting thing happening in the academic literature right now isn’t
discovery papers. It’s in the clinical and translational work on external controls,

federated learning, digital twins, and trial generalizability. These are the papers tend to get cited in regulatory guidance but not in Andreessen Horowitz market which is itself a signal worth paying attention to.

Start with external controls, which are quickly moving from statistical afterthou computable infrastructure. The FDA's 2025 draft guidance on externally control trials is notable not because it blesses the approach uncritically, but because of v it demands. External control and trial populations must be closely matched on baseline characteristics, disease severity, standard of care, prognostic factors, ou definitions, and follow-up procedures. That sounds like standard epidemiology, read it more carefully and what it describes is a substantial technical stack: pher normalization across source systems, covariate harmonization across sites and ti periods, temporal alignment to account for different eras of standard of care, endpoint ontology mapping, and robust causal adjustment for measured confour Building this reliably and in a way that satisfies an FDA statistician under adver review conditions is genuinely hard.

The more interesting development on the external control front is federated. A 2 Nature Communications paper introducing FedECA presents a federated learni approach for constructing external control arms with time-to-event outcomes in distributed data settings. The motivation is important: most of the data you'd wa use as external comparators lives in health systems that can't or won't share raw records. Privacy law, data use agreements, institutional politics – all of it creates friction that centralized data pooling approaches hit immediately. Federated me that can train and validate comparator models without moving raw data aren't ju technical curiosity; they're a necessary condition for scaling this infrastructure a the real-world data landscape.

The entrepreneurial implication here is underappreciated. The most valuable evidence company of the next cycle might not be a CRO or a data broker. It migh privacy-preserving cross-institution evidence network – something like a distrib proof system for regulatory comparators – that can assemble external control co without ever centralizing the underlying records. That's a different business mo than what most people are building.

On generalizability, a 2025 Nature Medicine study called TrialTranslator is one c cleaner demonstrations of the problem and the solution space. The authors shov real-world survival outcomes for oncology patients are often substantially worse RCT results, citing median overall survival reductions of roughly six months in s settings from prior work. They also find that about one in five real-world oncolo patients wouldn't qualify for a phase 3 trial, and they argue that eligibility criter alone don't explain the gap – selective recruitment of better-prognosis patients i trials is also a contributor. Their methodological response is to use ML-based prognostic phenotyping combined with trial emulation to test how well landmar RCTs generalize across different real-world risk strata.

The investment framing here is worth sitting with. If a trial result isn't a scalar - a distribution over subpopulations and care settings, some of which see somethi close to the trial effect and some of which see much less – then the ability to systematically quantify where a drug's efficacy attenuates in the real world is genuinely valuable. It changes launch forecasting. It changes payer strategy. It ch medical affairs planning. It potentially changes M&A valuation in situations whe sponsor is trying to figure out how much of a phase 3 result will survive contact real-world prescribing patterns. Trial informatics reframed as asset diligence infrastructure is a very different market than trial informatics reframed as work automation.

Digital twins are the most overhyped piece of this stack, and also the one that's r consequential if the validation problem gets solved. The marketing use of "digit twin" in health and life sciences covers a huge range of things, most of which are twins in any meaningful modeling sense. But the academic literature is getting r specific. A 2025 npj Systems Biology and Applications paper argues that properl constructed twins can improve trial efficiency, enable early adverse event detecti and contribute to adaptive trial design – while also flagging serious ethical and validation concerns that remain unresolved. A translational example cited in the npj Drug Discovery review involved systems-based digital twins in a Phase I mosunetuzumab study in NHL to characterize dose-response and identify predic

biomarkers, which then informed adaptive dosing protocols. That's a concrete cl
application, not a marketing slide.

The catch, which any serious investor should dwell on, is that twin technology is
investable at scale if verification, validation, and uncertainty quantification (VV
matures. A 2025 npj Digital Medicine survey on this topic makes the point clear
VVUQ is not optional infrastructure for a twin that's going to support safety- ar
efficacy-relevant decisions in a regulated environment. It's the central requireme
The moat in digital twins isn't the simulation demo. It's the validation framewor
uncertainty propagation methodology, and the decision thresholding protocol th
makes the model usable under regulatory scrutiny. That's a much more defensib
much harder thing to build, and it's not what most current twin vendors are sell

On recruitment and patient matching, the TrialGPT paper published in Nature
Communications in 2024 introduced an end-to-end LLM framework for zero-sh
patient-to-trial matching. Subsequent 2025 and 2026 work has pushed into
multimodal inputs, ingesting structured EHR data alongside clinical notes and c
chart artifacts. Progress is real. But the subtler argument is that recruitment alo
probably too narrow a moat for a durable infrastructure business unless it expar
into a broader trial-state architecture: protocol criteria management, eligibility
evidence tracking, missing data handling, site capability assessment, patient
longitudinal state representation, and expected screen-failure risk prediction. T
companies that capture enrollment as an entry point and then build upward intc
broader state graph are the ones worth watching.

A 2025 npj Digital Medicine scoping review that examined 142 AI-in-trials stud
from 2013 through 2024 found growing AI use across safety monitoring (55 stud
efficacy assessment (46), and operational risk management (45). But the same rev
makes clear that most of this work is retrospective, many datasets are
nonrepresentative, and prospective reproducible evidence for these tools is still
limited. That's both a problem and an opportunity. The winning businesses here
be the ones with the strongest benchmark on a retrospective dataset. They'll be
ones that have figured out prospective deployment, label scarcity, site workflow
integration, and auditability under regulatory review. Those are organizational a

technical challenges simultaneously, which is exactly the kind of problem that's to replicate quickly.

Finally, on platform and adaptive trial architecture, a 2024 Signal Transduction Targeted Therapy review makes the case that basket, umbrella, and platform tri designs are increasingly aligned with the biological logic of precision medicine a multi-omics and sequencing make molecular stratification more granular. Platfo trials can dynamically add and drop arms, incorporate biomarker-guided stratification, and use interim evidence to adapt allocation rules. Once the trial architecture becomes this dynamic and biomarker-dependent, software isn't bac office tooling anymore. It becomes part of the protocol logic itself. That's a diffe market than generic eClinical infrastructure, and it's more likely to support businesses with real defensibility.

Regulators are opening the door but raising the bar

The FDA's posture over the last two years is worth reading carefully because it's misread in two opposite directions simultaneously. One camp thinks the agency becoming more conservative and therefore all this novel methodology work is speculative. The other camp thinks the agency is embracing new tools enthusias and that regulatory friction is going away. Both camps are wrong.

What the agency is actually doing is more nuanced: it's conditionally opening th to external controls, real-world data, digital health technologies, and model-assi endpoints while simultaneously raising the evidentiary bar for how those tools n be validated and deployed. The 2025 externally controlled trials draft guidance i clearest example. The agency isn't saying don't use external comparators. It's say here is the technical specification for what a regulatory-grade external comparat actually requires, and that specification is demanding. The same pattern shows u FDA guidance on digital health technologies for drug development and in the framework for AI and machine learning in drug development – conditional open

combined with explicit expectations around transparency, validation, and humar oversight.

That combination is exactly what creates a wedge for infrastructure companies. a regulator says "we will accept this approach if you can demonstrate X, Y, and Z and X, Y, and Z are technically hard and require significant organizational capab to demonstrate, then whoever builds the infrastructure to demonstrate X, Y, and reliably becomes a necessary part of every program that wants to use these approaches. That's a strong structural position.

The European Medicines Agency has been running parallel threads on adaptive and real-world evidence that broadly reinforce this direction, and the Internatio Council for Harmonisation's E6(R3) update to GCP guidance signals a similar m toward risk-based monitoring and more flexible trial architectures. The regulato environment isn't creating a free pass. It's creating a technical specification for a generation of development infrastructure, and that specification is detailed enou that building to it is a real engineering problem.

What this means for founders, CROs, a venture underwriting

The practical implications of this analysis split across three groups: founders de what to build, CROs thinking about where their business model goes, and VCs figuring out where the durable value accumulates.

For founders, the clearest takeaway is to think carefully about which part of the evidence stack you're building and whether it's a wedge or a destination. Recruit is a real problem and a reasonable entry point, but it's also the most crowded an most likely to get commoditized by EHR vendors and large CROs with existing s relationships. The more durable positions are likely in the layers that sit beneath beside recruitment: comparator construction, phenotype infrastructure, continu measurement, model validation, and adaptive protocol software. These are harde build, slower to sell, and less intuitive to demo, but they also sit closer to the act

evidence package that a sponsor needs to file, which means they're harder to cut the budget when things get tight.

The federated architecture point deserves particular emphasis for founders. The that would make the best external comparators – rich longitudinal records from health systems covering diverse patient populations – is almost never going to be centralizable in a single commercial database. The legal, political, and competiti dynamics are too complicated. The companies that figure out how to generate regulatory-grade evidence from distributed data without pooling raw records are building something that addresses a structural constraint rather than a workflow inefficiency, and structural constraints tend to create better businesses.

For CROs, this is both an opportunity and an existential threat depending on ho they respond. The large CROs have site relationships, operational infrastructure regulatory track records that are genuinely hard to replicate. But their core competency – running trials as defined – is increasingly being challenged by the question of whether trials as currently defined are the right design in the first pl adaptive master protocols with dynamic arm allocation and biomarker-driven enrollment become more common, the operational model of the large CROs need change significantly. The CROs that are building or acquiring the computationa infrastructure for external controls, digital endpoints, and adaptive design are positioning well. The ones that are treating software as a commodity layer they c outsource are making a mistake.

For VCs, the most important framing shift is to stop thinking about this market for trials" and start thinking about it as evidence infrastructure, which is more analogous to financial market data infrastructure than to SaaS workflow tools. T businesses with the highest long-run value in financial data didn't build fancy analytics. They built the underlying data network – the feeds, the normalization audit trail, the compliance layer – that every analytics product depended on. The analogous businesses in clinical development will be those that own the provena chain from raw clinical and real-world data to regulatory-grade evidence submis That chain involves a lot of unglamorous infrastructure: source data verification

protocol versioning, endpoint adjudication documentation, uncertainty reportin
human override tracking. None of it is conference-talk material. All of it is nece

The five infrastructure layers that matt

To make this concrete, here's where the category-defining businesses of the next
development cycle are most likely to form.

The first layer is comparator infrastructure: external controls, target trial emula
and federated evidence networks. This is arguably the most technically demandi
and the most immediately actionable given the FDA's draft guidance providing a
explicit requirements specification. The players that build federated comparator
networks with real governance, real privacy architecture, and real regulatory tra
records will own a capability that every sponsor running a single-arm trial or a s
randomized study needs access to.

The second layer is phenotype infrastructure: multimodal risk stratification,
biomarker enrichment, and generalizability scoring. The TrialTranslator work p
to this directly. If trial results are distributions over subpopulations rather than
scalars, the ability to characterize those distributions systematically – using ML-
phenotyping across structured and unstructured data – creates value at multiple
stages of development and commercialization. This layer is also closely connecte
the precision medicine thesis, since the biomarker enrichment work that makes
platform trials more efficient is the same capability that makes generalizability
assessment more rigorous.

The third layer is continuous measurement infrastructure: digital health techno
remote endpoint capture, and sensor-grade clinical measures. The FDA has beer
investing significant guidance bandwidth in digital health technologies for drug
development, and the clinical community has been developing validation framev
for patient-reported outcomes and wearable-derived endpoints. The infrastructu
play here is less about the device itself and more about the data architecture,
validation documentation, and endpoint certification pathway that turns a devic
generated signal into a regulatory-grade clinical endpoint.

The fourth layer is model assurance infrastructure: verification, validation, uncertainty quantification, and auditability for twins and predictive models. Thi currently the most underbuilt part of the stack. Lots of companies have models. few have the frameworks to demonstrate that their models behave reliably unde distribution shift, can propagate uncertainty in ways that support decision thresholding, and maintain an audit trail that satisfies regulatory review. That ga the bottleneck for digital twins specifically but also for any predictive model tha wants to be used in a regulated decision context.

The fifth layer is adaptive protocol infrastructure: software for master protocols, dynamic allocation, biomarker-driven arm management, and interim evidence integration. As the trial architecture itself becomes more computationally native software that manages the protocol logic – not just the eClinical workflow but th actual decision rules for arm allocation, stopping, and enrichment – becomes cri path. This is still early, but the academic momentum behind basket, umbrella, a platform designs combined with the FDA's growing comfort with adaptive desig suggests the market is building.

The unified thesis across all five layers is the same: AI is industrializing hypothe generation faster than the industry can industrialize evidence generation. The molecule problem, while not fully solved, is no longer the primary constraint. Th evidence problem – who can assemble trustworthy, regulatory-grade, causally defensible proofs that a drug does what it's supposed to do in patients who actu exist – is the constraint now. That's where the next generation of durable health- infrastructure companies will be built, and the window to form them is open.

5 Likes • 1 Restack

← Previous

Discussion about this post

Comments Restacks

Write a comment...

© 2026 Thoughts on Healthcare · Privacy • Terms • Collection notice
Substack is the home for great culture