sapien labs

Faster, cheaper clinical trials.

We create digital twins using real medical data so drug makers recruit fewer patients while reaching the same result.

1M+
de-identified patient records

550k+
clinician notes structured

$500k
Raised till date from Lightspeed and EF

Problem

Clinical Trials are following EROOM's Law

A late-stage trial costs over **$50,000** per patient and needs about **2,000** of them. **Half** are given a sugar pill (the control group), only to set a baseline the real drug is measured against.

Sponsors have tried to reuse medical datasets they bought at high prices to build this baseline instead. That data fails to meet the FDA's quality bar.

> *"We have bought datasets for over $70M for other studies. For this, they are practically useless."*
> Director of Biostatistics, global pharma company

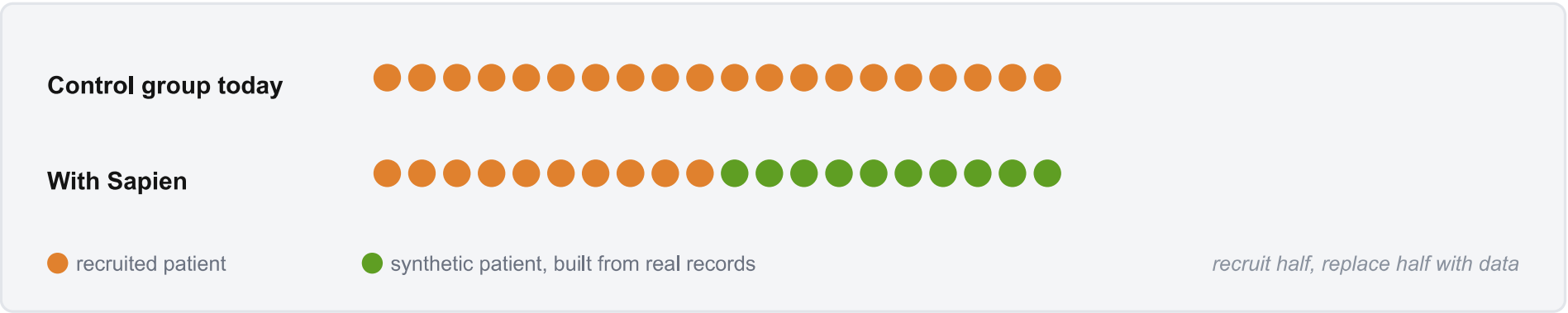
$50k+
cost per patient recruited

37%
of trial sites fail to enroll enough patients

11%
of sites enroll no one at all

Solution

We rebuild most of the control group from real medical records.



We hold over a million real patient records. From these we build many control groups, and each sponsor selects the subset that fits their trial. A US medtech is already generating trial evidence with us at a fraction of the usual cost.

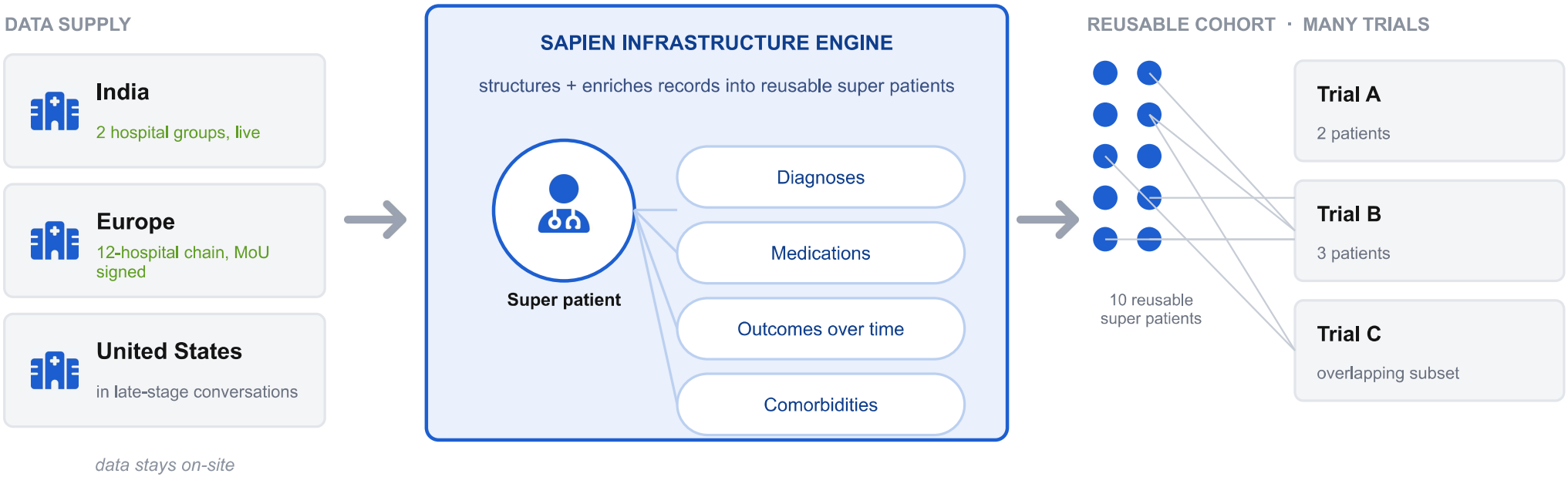
$40k+
saved per patient

50%
of the control group replaced

10x
lower cost

Product

One data engine. Every trial draws from it.



Trial revenue funds the structuring on-site and is shared back with partner hospitals.

Why Now

FDA is opening doors themselves and we found the sweet spot!



Objective-data players entered years ago: **Flatiron** (oncology, 2012), **Unlearn** (neurology, 2017). The subjective side stayed untouched.

The 2030 patent cliff

Around $170B of drug sales lose patent protection by the early 2030s. That has doubled biotech investment and the number of drugs racing through trials.

The technology is finally ready

Clinical records are digitized at massive scale, and modern AI models can now structure the written, subjective parts of psychiatric care that older tools could not.

Competition

Everyone is shrinking the control group with data. We own the corner they avoided.

They differ on two things: whether they use **objective data** like lab values and scans, or the **subjective data** in clinical notes, and whether their patients are **reusable** or rebuilt from scratch each time.

We work with the subjective data others avoided, and our cohorts are reusable, so our cost falls with every trial.



Market

The trial market is a ladder. We enter at the bottom and climb.

Big Pharma
biggest budgets, buys on time saved

Mid Pharma
large trials, slower cycles

Biotech (Series B+)
first paid pilots in reach

Digital therapeutics & medtech
the placebo problem is worst here

Academic labs
40 to 50 patients, one lead researcher

we start here and climb up ↑

~$1B	SOM — biotech & medtech beachhead
$5B	SAM — big pharma at 50% replacement
~$40B	TAM — psychiatry

Bottom-up: $70M trial × 25% of patients replaced = ~$17M saved. Our 10% = $1.7M per program, across ~300 programs and 2 phases ≈ $1B.

Research

arXiv

Search Submit Donate Log in

Statistics > Applications

[Submitted on 7 Jul 2026]

Psych-ECA: A Reproducible Semi-Synthetic Benchmark for Synthetic Control Arms in Longitudinal Psychiatry

Aakash Bhagat, Shashank Choudhary

External and synthetic control arms (ECAs) are entering psychiatric drug development, but the field lacks a benchmark that evaluates the properties regulators care about: not only how accurately a method reconstructs untreated trajectories, but whether its uncertainty is calibrated, whether it is robust to the informative observation times common in mental-health records (sicker patients are seen more often), and what false-positive rate it induces in go/no-go trial decisions. Real psychiatric trial data (e.g. STAR-D and registry cohorts) require credentialed access and lack ground-truth counterfactuals, so, following established semi-synthetic benchmarks in causal inference (IHDP, ACIC, and the PK-PD tumor-growth simulator), we release Psych-ECA, a fully reproducible generator of longitudinal symptom trajectories for depression (PHQ-9), anxiety (HAM-A), and psychosis (PANSS) with known counterfactual control arms, informative visits, and validated-scale measurement noise. We benchmark eight estimators spanning carry-forward, pooled real-world-data averages, nearest-neighbour matching, linear mixed models, gradient boosting, and the Scribe trajectory-bridge method. Three findings emerge. First, trajectory and flexible machine learning methods achieve the best counterfactual accuracy (about 2.3 PHQ-9 RMSE), outperforming cross-sectional baselines. Second, only Scribe is both accurate and calibrated, achieving 93-96% empirical coverage of nominal 90% prediction intervals, compared with 87-88% for gradient boosting and 62-75% for uncalibrated SDE models. Third, inverse-intensity correction reduces bias under informative sampling, while Scribe's calibrated intervals are the only trajectory method that maintains nominal false-positive rates as informativeness increases. We release all code, data-generation scripts, and random seeds to enable fully reproducible evaluation.

Subjects: **Applications (stat.AP)**; Machine Learning (cs.LG)
Cite as: arXiv:2607.27224 **[stat.AP]**
(or arXiv:2607.27224v1 **[stat.AP]** for this version)
https://doi.org/10.48550/arXiv.2607.27224

Submission history

From: Aakash Bhagat [view email]
[v1] Tue, 7 Jul 2026 16:04:16 UTC (49 KB)

Access Paper:
View PDF
HTML (experimental)
TeX Source
view license

Current browse context:
stat.AP
< prev | next >
new | recent | 2026-07
Change to browse by:
cs
cs.LG
stat

References & Citations
NASA ADS
Google Scholar
Semantic Scholar

Export BibTeX Citation

Bookmark

Bibliographic Tools | Code, Data, Media | Demos | Related Papers | About arXivLabs

Bibliographic and Citation Tools

Team

A young Bio x ML team, backed by patient capital and senior advisors.



COFOUNDER & CEO

Shashank

IIT Delhi, Biotechnology

Background in gene therapy and microfluidics



COFOUNDER & CTO

Aakash Bhagat

BITS Pilani, Electrical + Maths

Built ML models for Turing-award labs in the US

ADVISORS

Advisors affiliated with Pfizer, GSK, Tufts, Johns Hopkins, Imperial, and UCL.

BACKED BY

Lightspeed · Entrepreneur First · Transpose Platform

$500k+ raised to date + $300k committed

The Ask

Raising a $2M pre-seed on a SAFE.

$500k+ in **$2M target**

Raised so far from Lightspeed, Entrepreneur First, and Transpose Platform.

- A Delaware C-corporation with operating subsidiaries in India for delivery.
- This round takes our existing paid studies into FDA-grade trials and wins our first big-pharma program within 12 months.
- It funds the infrastructure to process over 5 million patient records in the next year.

$500k
revenue in 6 months

$2M
revenue in 12 months

shashank@sapienlabs.bio · sapienlabs.bio

Startup: Sapien Labs

https://iitstartups.org/us-demoday-cohort-15/

Sapien Labs collects and enriches clinical data to construct patient digital twins that replace human subjects in Phase 2/3 clinical trials, saving cost and time for pharma.

SCAN

https://zfrmz.com/p19jyuetieRIM6wRPsfF



Shashank Choudhary

Co-Founder & CEO

shashank@sapienlabs.bio

+1 617 402 2226

https://www.linkedin.com/in/heyshashank/

https://cal.com/shashanksapien/us