

The Health System Data Monetization Cartel: Why the Most Valuable Dataset in Life Sciences Is Sitting on the Table

FEB 21, 2026



4



1

Share

Abstract

The structural case for a for-profit clinical data cooperative that actually captures economic value of real-world health system data for pharma, biotech, and AI model training markets.

Key claims:

- Real-world clinical data (structured EHR, imaging, genomics, pathology) is worth orders of magnitude more than health systems currently extract from it
- Existing cooperative models (Truveta, TriNetX) leave massive money on the table through timid pricing, weak IP posture, and misaligned incentive structures
- A coalition of 20 major health systems with 10-20M longitudinal patient records could generate \$500M+ annually in pharma data licensing alone
- The AI model training market creates an entirely new and arguably larger revenue stream that existing players are barely touching
- Health systems as equity holders, not just participants, is the structural key that changes everything

Data points referenced:

- Global RWD/RWE market: \$2.5B in 2023, projected \$4.8B by 2028 (CAGR ~14%)

- Pharma spends est. \$3-5B annually on synthetic data, claims proxies, and limited real-world datasets
- Clinical trial recruitment failures cost the industry \$8B+ annually
- Top foundation model companies (OpenAI, Google DeepMind, Mistral, etc.) have scalable access to structured clinical data
- Truveta raised \$200M at a valuation that implies significant underpricing of the underlying asset

Table of Contents

The Setup: What Real-World Clinical Data Actually Is

The Market Failure Nobody Is Fixing

What the Existing Players Got Wrong

The Cartel Structure: How to Actually Build This

The Revenue Stack

The Operational Reality

Why Now

The Exit

The Setup: What Real-World Clinical Data Actually Is

To understand the opportunity, it helps to be precise about what “clinical data” because the term gets thrown around in health tech circles to mean basically everything and therefore nothing. Claims data is not clinical data. Survey data is clinical data. Patient-reported outcomes from a wellness app are definitely not c

data. What the life sciences industry actually needs, and mostly cannot get at scale, is structured longitudinal records from electronic health systems including problem lists, medications, labs, vital signs, procedures, imaging reports, pathology results, and increasingly genomic data all tied together at the patient level over time.

This is what gets generated every day in health systems across the country and largely disappears into archive storage never to be touched again except for billing purposes. The clinical encounter generates a staggering volume of information: a typical hospitalization might involve dozens of lab values, imaging reads, nursing assessments, physician notes, medication administration records, and procedure codes. A patient with a chronic condition like diabetes or heart failure accumulates years of longitudinal data points across outpatient visits, hospitalizations, specialist consults, and pharmacy interactions. Multiply that by millions of patients across a major health system and the dataset is enormous by any reasonable definition.

The structured portion of this, meaning the data that is actually in discrete fields rather than buried in free text notes, is particularly valuable. Lab values with reference ranges and timestamps. Medication lists with dosing and duration. Diagnosis codes mapped to encounter dates. Vital signs in time series format. This is the stuff that actually moves drug development forward because it can be queried, analyzed, and modeled without massive natural language processing overhead. Imaging and pathology data adds another layer entirely because you now have radiologic diagnostic content tied to outcomes in a way that is genuinely impossible to replicate with synthetic approaches.

The genomic layer is where things get really interesting and where the long-term value of the asset class becomes clearer. Health systems that have implemented biobanking programs, and there are more of them than most people realize, are on germline and somatic genomic data tied to phenotypic clinical records in a way that pharmaceutical companies would pay almost anything to access at scale. The UK Biobank demonstrated what this kind of linked genomic-clinical dataset is worth to the research community and it was built with public funding and essentially given away for free. The American version of that asset, built as a for-profit entity, would look very different economically.

The Market Failure Nobody Is Fixing

Here is the fundamental problem. Health systems generate this data as a byproduct of patient care, pay significant money to store and manage it, and then extract almost no economic value from it beyond their core clinical and billing operations. The occasional academic research collaboration generates nominal grant overhead. If approved data sharing arrangements with pharma sponsors are often structured as cost-recovery deals that barely cover the administrative burden of data preparation. The idea that this data has independent commercial value and that health systems should be capturing that value aggressively is genuinely foreign to most health system leadership teams.

Meanwhile pharma and biotech are doing increasingly acrobatic things to approximate the clinical insight they cannot get from real-world sources. Claims data has been the dominant proxy for years and the industry has spent enormous resources building sophisticated analytics on top of Medicare and commercial claims to infer things like disease progression, treatment patterns, and outcomes. The fundamental limitation is that claims capture billing events, not clinical reality. A claim tells you that a patient had an office visit coded as a diabetes management encounter. It does not tell you what their HbA1c was, whether they were adherent to their medication, what their comorbidity burden looked like in clinical detail, or how their condition actually progressed over time. The gap between what pharma needs and what claims can provide is wide enough to drive a truck through.

Synthetic data has become a fashionable workaround and some genuinely impressive technical work has been done on generative approaches to clinical data synthesis. An honest assessment is that synthetic data is useful for software development, algorithm testing, and certain types of statistical modeling but it has fundamental limitations for anything requiring authentic population-level signal. You cannot synthesize a pharmacovigilance signal. You cannot train a clinical AI model on synthetic data and expect it to generalize to real patient populations. You definitely cannot use synthetic data for regulatory submissions where FDA expects real-world evidence.

The total spend across claims data vendors, synthetic data companies, limited real world data licenses, and related infrastructure is in the \$3 to \$5 billion annual range and growing fast. None of this money is going to the health systems that actually own the underlying data. It is going to intermediaries who have figured out how to package and resell inferior proxies because the real thing was not organized or available at scale. This is the market failure in one sentence: the people who own the best asset are not participating in the market for it.

What the Existing Players Got Wrong

Truveta and TriNetX are the two most visible attempts to build something like a health system data cooperative and both of them, for different reasons, illustrate exactly the mistakes to avoid if the goal is to actually capture the economic value of the underlying asset.

Truveta, which raised around \$200M from a group of major health systems including Providence, CommonSpirit, Ascension, and others, is technically impressive. The infrastructure is real and the governance model was thoughtful. The problem is structural and pricing-related. Truveta was built with a cooperative ethos that prioritized broad access and research enablement over aggressive value capture. The pricing reflects this. Academic and nonprofit customers get favorable terms. Pharmaceutical pricing, while not publicly disclosed, is understood in the industry to be well below what the underlying asset would support in a purely commercial pricing environment. The health system members receive shares in a company whose valuation, given its revenue and pricing strategy, significantly undervalues the data asset they contributed.

More importantly, Truveta was designed from day one to be a cooperative infrastructure company rather than a commercial data business. That sounds like a subtle distinction but it is not. A cooperative infrastructure company optimizes for breadth of access and community benefit. A commercial data business optimizes for revenue per record and margin per transaction. These are not the same objective

function and you cannot serve both simultaneously without compromising on the commercial side, which is exactly what happened.

TriNetX is a different model and a different set of problems. The company operates a network facilitator that allows pharma sponsors to query across a distributed network of health system databases for clinical trial feasibility and recruitment purposes. The health systems in the network are essentially providing a service for modest or no compensation in exchange for being connected to clinical trial opportunities. The value exchange is extremely lopsided in favor of pharma, and health systems participate because trial sponsorship revenue is something they understand and value, not because they have thought clearly about the standalone value of their data asset.

Neither model contemplates what is arguably the most important structural principle: health systems should not just be members or participants. They should be equity holders in a for-profit entity whose explicit mission is to maximize the commercial value of the data asset they collectively own. The difference between contributing to a cooperative in exchange for governance rights and owning equity in a company that is aggressively monetizing your data contribution is enormous when you run the numbers forward five to ten years.

There is also a pricing posture problem with both existing players that reflects a misunderstanding of negotiating leverage. A coalition of 20 major health systems with 10 to 20 million unique longitudinal patient records has a genuinely monopolist position in the market for high-quality real-world clinical data at scale. Pharma companies do not have good alternatives. They are price-sensitive but not infinitely so, and the value they derive from accessing high-quality clinical data at scale for drug development and pharmacovigilance purposes is orders of magnitude greater than what the existing players charge. The current pricing paradigm reflects the supply side's underestimation of its own leverage, not any fundamental constraint on what the market would bear.

The Cartel Structure: How to Actually Build This

The name matters more than it might seem. Calling this a cartel is intentional and accurate. A cartel is a group of independent entities that coordinate to control the supply and pricing of a commodity in ways that maximize collective return. That is precisely the structure that the clinical data market needs and that health systems have every right to build. The legal framework for this kind of coordination among healthcare entities is well-developed, the antitrust exposure is manageable if the entity is properly structured, and the precedents from other data consortium models in financial services and telecommunications are instructive.

The entity structure should be a for-profit C-corp with health systems as founding equity holders. Not a cooperative, not a nonprofit, not a joint venture with a data vendor as the operating partner. A genuine commercial enterprise where the health system equity stake is proportional to data contribution measured in attributed patient lives, record completeness, and longitudinal depth. This alignment mechanism is critical. When health systems own equity that appreciates with revenue, they have an incentive to contribute their best data, maintain quality, and advocate for aggressive pricing in ways that cooperative participants simply do not.

The founding coalition matters enormously and the target should be a set of systems that between them represent geographic diversity, patient population diversity, and depth of clinical data capture. Twenty systems is the right order of magnitude. You want enough patient lives to be statistically meaningful for rare disease research, enough geographic spread to avoid regional sampling bias, and enough health system diversity to include both academic medical centers with research infrastructure and large community systems with high patient volumes. The ideal founding coalition probably includes three or four academic medical centers that bring the credibility and research infrastructure, a similar number of large regional systems that bring volume, and a mix of specialty-focused systems that bring depth in specific therapeutic areas.

The governance model needs to be commercial-grade rather than academic-grade. This is where cooperative models typically fail. Academic and nonprofit governance structures are optimized for consensus, equity, and stakeholder representation. They are not optimized for commercial decision-making speed, pricing discipline, or aggressive market positioning. The board of the entity should include health system representation alongside genuine commercial operators who understand data licensing, pharma procurement, and technology pricing. The CEO should come from commercial health tech or data licensing, not from academic medicine or health system administration.

Data standardization and curation is the operational core of the business and deserves more attention than it usually gets in the strategic discussion. The raw data that health systems generate is not a commercial product. It is a collection of EHR exports in various formats with varying degrees of structure, completeness, and accuracy. Turning that into a queryable, analytically-ready dataset that pharma and AI customers can actually use requires significant ongoing investment in data engineering, terminology standardization, quality assurance, and de-identification infrastructure. This is not a one-time transformation project. It is a continuous operational function that requires real technical capability. The governance model needs to allocate meaningful budget to this function and the founding documents need to require health system participation in data quality improvement as a condition of equity maintenance.

The Revenue Stack

The base revenue layer is pharma data licensing and this alone justifies the business model. Pharmaceutical companies use real-world clinical data across several high-value use cases: regulatory submissions requiring real-world evidence for label expansions and post-marketing commitments, pharmacovigilance and safety monitoring that FDA increasingly requires as a condition of approval, comparative effectiveness research that informs formulary positioning and payer negotiations, patient identification for clinical trial recruitment and feasibility analysis.

Each of these use cases has a different willingness-to-pay profile and a different purchase decision structure, but all of them involve material budget allocations for major pharma companies. A conservative estimate of what a coalition of 20 health systems with 15 million longitudinal patient records could extract from pharma licensing is \$200 to \$300 million annually at current market pricing. That is the conservative case. At pricing that actually reflects the leverage position of controlling the supply of best-in-class real-world clinical data, the number is \$400 to \$600 million. These are not speculative figures. They are derived from known per-patient-year pricing for premium real-world data assets and known pharma spending patterns on real-world evidence.

Clinical trial recruitment is a separate and arguably more defensible revenue stream. The cost of failed clinical trial recruitment is staggering. The industry average fully loaded cost of a Phase 3 recruitment failure, including protocol amendments, trial extensions, and lost development time, runs into nine figures for large trials. The value proposition of being able to identify and pre-screen patients who meet trial inclusion criteria based on actual clinical records rather than claims approximation is massive and the pricing should reflect it. A per-patient-identified fee structure for trial recruitment assistance, combined with site feasibility fees for the health systems that host the recruited patients, creates a revenue model that is aligned with value delivery in an unusually clear way.

The AI model training market is the revenue stream that existing players are almost entirely ignoring and it may ultimately be larger than the pharma licensing business. Every major foundation model company is acutely aware that their next generation of clinically capable models is bottlenecked by access to high-quality, structured, real-world clinical data at scale. OpenAI, Google DeepMind, Microsoft/Nuance, and other serious clinical AI startups have the same problem: they can access enormous quantities of medical literature, synthetic data, and curated public datasets but they cannot access the authentic clinical records that would allow their models to generalize to real patient populations in real clinical settings.

The contract structures for AI model training are different from pharma licensing. Rather than per-patient or per-record pricing, AI model training deals are typically

structured as large upfront payments for specific training runs plus ongoing access fees for model fine-tuning and evaluation. The deals that exist in adjacent data categories suggest this market could generate \$50 to \$150 million annually for a coalition-scale clinical dataset, with significant upside as model capabilities advance and demand increases. The strategic value of being the primary training data provider for the leading clinical AI models also creates durable competitive positioning that compounds over time in ways that transactional pharma licensing does not.

Insurance and payer analytics represents a third revenue layer that is smaller but strategically valuable. Payers have chronic problems with clinical data access that is structurally similar to pharma's problems. They use claims as a proxy for clinical reality when making coverage decisions, care management interventions, and risk stratification determinations. Access to actual clinical records for their attributed populations would improve all of these functions materially, and the payer industry has demonstrated willingness to pay for data assets that improve clinical and actuarial performance. This market is probably \$50 to \$100 million annually at scale and provides diversification away from pharma as the primary customer concentration.

The combined revenue potential across pharma licensing, clinical trial services, model training, and payer analytics is \$600 million to \$1 billion annually for a medium coalition-scale entity. The margin profile is exceptional because the marginal cost of additional data licensing is low once the core data infrastructure is built. Software-like gross margins, probably 60 to 70 percent at scale, on a revenue base of this size produces an EBITDA profile that justifies a valuation in the multi-billion dollar range even at conservative multiples.

The Operational Reality

The hard part of building this is not the business model or the financial projections. The hard part is health system alignment and the organizational complexity of coordinating 20 large, bureaucratic, legally conservative institutions around a commercial objective. This is where most coalition-based health tech ventures die.

from market failure but from governance failure and principal-agent problems within the founding group.

Health systems move slowly for reasons that are structural, not incompetent. They have legal and compliance functions that are appropriately cautious about novel arrangements. They have governance processes that require board approval for significant business decisions. They have strategic priorities that are almost entirely focused on clinical operations and financial sustainability rather than data commercialization. And they have constituencies, including medical staffs, patient advocacy groups, and community stakeholders, who have legitimate questions about how their data is being used even in de-identified form.

Navigating this requires a different approach than typical enterprise sales or partnership development. The founding CEO of this entity needs to be someone who understands health system governance at a deep level and has existing relationships with C-suite leadership at major systems. Former health system executives who have also operated in commercial health tech are the rare profile that works here. The mover advantage in getting founding commitments from a credible coalition of 2-3 systems is enormous and the barrier to replication once that coalition is assembled is extremely high. Getting there requires sustained relationship-based work over a 18 month period that cannot be shortcut.

Data governance and patient consent frameworks are real operational challenges that deserve serious attention rather than hand-waving. The HIPAA framework for de-identified data use is well-established and a properly structured data use architecture can operate within it, but the details matter and the reputational risk of getting wrong is existential. Some patient advocates will oppose any commercial data use regardless of de-identification approach, and the political environment around health data privacy has become more complex over the past several years. The entity needs to invest in a genuine patient trust framework, including transparent opt-out mechanisms, clear public communication about data use, and ongoing engagement with patient advocacy communities, not as a PR exercise but as a genuine operational commitment.

Regulatory positioning matters too and is frequently underweighted by data founders who think of FDA as primarily relevant to drug and device approval rather than data commercialization. The way data is prepared, documented, and delivered for regulatory submission purposes affects how FDA views real-world evidence from this data source, and FDA's endorsement or skepticism of a particular RWD source has enormous commercial implications for pharma customers. Building a regulatory affairs capability that proactively engages with FDA on real-world data standards, participates in relevant pilot programs, and develops a track record of FDA-accepted RWE studies is a multi-year investment that pays off in the form of premium pricing and reduced commercial risk for pharma customers.

Why Now

The timing argument for this is stronger than it has been at any point in the past decade and probably as strong as it is going to get for the next several years. Several converging factors create a window that is real but not permanent.

The AI model training demand is a new and genuinely time-sensitive component. Foundation model companies are making major bets right now on clinical AI capabilities and the data access strategies they establish in the next 18 to 24 months will shape the competitive landscape for clinical AI for years. Being the primary training data provider for the leading clinical AI model is a strategic position worth capturing aggressively. Two years from now the first-mover advantage in this space will be significantly smaller because the field will have consolidated around a small number of solutions with established data partnerships.

Health system financial pressure is creating receptivity to revenue diversification that did not exist five years ago. The post-COVID financial environment for health systems has been consistently challenging with labor cost inflation, payer mix pressure, and reimbursement rate constraints creating persistent margin compression at even well-managed systems. CFOs who previously would have dismissed data commercialization as a distraction are now genuinely interested in discussions about new revenue streams. This creates a brief window where the conversation is easier than it has

historically been, before either the financial environment improves or the health system strategy community converges on a consensus approach and the leverage position of early organizers disappears.

FDA's increasing requirements for real-world evidence create a structural demand driver that is regulatory rather than discretionary. The number of drug applications where FDA is requiring post-market real-world evidence as a condition of approval is growing, and the quality bar for that evidence is rising. Pharma companies that rely on claims-based RWE for earlier regulatory submissions are finding that FDA is increasingly skeptical of claims as a clinical proxy and more receptive to evidence derived from structured EHR data. This is a regulatory tailwind that creates durable demand for the core product.

The competitive landscape is not going to stay empty. Truveta exists, is capitalizing and is continuing to build its infrastructure. TriNetX is active and growing its network. Epic, which controls more EHR data than anyone else in the country, has an interest in this market that it has not fully expressed yet. The window for establishing a well-capitalized, aggressively-structured alternative with superior commercial alignment is probably two to three years before the competitive dynamics shift materially.

The Exit

The exit thesis for this business is unusually clear, which is not always true in health tech where acquirer interest is often speculative or dependent on business model pivots. Three credible exit paths exist and they are not mutually exclusive.

Strategic acquisition by a major pharma company or pharma services conglomerate is the most obvious path. IQVIA, Veeva, Symphony Health, and the pharma data divisions of major conglomerates would all have strategic interest in acquiring a coalition-scale clinical dataset with durable health system relationships and ongoing data supply. The valuation upside here is substantial because strategic acquirers would pay not just for the current revenue stream but for the competitive moat that comes with owning the supply relationship. This is the path that generates the largest

absolute return but also requires the most negotiating sophistication because the likely acquirers are sophisticated buyers with significant leverage.

IPO is viable at scale and the public markets have demonstrated appetite for health data companies with strong recurring revenue and defensible market positions. A comparable set of public health data companies trades at revenue multiples that would imply a multi-billion dollar valuation for a company generating \$600 million plus in annual revenue at software-like margins. The IPO path also serves the health system equity holders well because it creates liquidity without requiring them to sell a strategically important asset entirely.

The third path, staying private and distributing cash, is underrated in a sector that reflexively assumes venture-style exit is the only success criterion. A business generating \$600 million in revenue at 65 percent gross margins produces enough cash flow to distribute material returns to health system equity holders annually while continuing to invest in infrastructure and capability. Not every valuable business needs to be sold or taken public and the health system equity holders, who are not venture funds with mandatory return timelines, might actually prefer a durable cash-generating asset to a liquidity event that terminates their participation.

The total value creation opportunity here, across the pharma licensing market, AI training contracts, and payer analytics, is large enough to justify describing it as a once-in-a-generation asset formation opportunity in health tech. The underlying commodity, authenticated longitudinal clinical records at scale, is genuinely irreplaceable. The market failure is real and well-documented. The structural solution is clear even if the execution is hard. What is missing is not insight into the opportunity but rather the commercial conviction and organizational capability to build the entity that captures it before the window closes.



4 Likes • 1 Restack

← Previous

Next →

Discussion about this post

Comments

Restacks



Write a comment...