Mining Gold from Public Health Data: The Unrealized Opportunity in Healthcare’s Open Datasets

NOV 10, 2025 • PAID

Disclaimer: The views and opinions expressed in this essay are my own and do not refl views of my employer, Datavant, or any other organization with which I am affiliated.

Table of Contents

Abstract

This essay examines the emerging landscape of public and private healthcare dat exploring viable business models for health tech entrepreneurs and angel investo We analyze ten major data resources spanning both freely accessible and comme datasets: Google Cloud Healthcare API, CDC Open Data, HHS/[HealthData.gov] (http://HealthData.gov) portals, SatHealth multimodal environmental health data SDOH-NLI clinical notes corpus, MIMIC-IV and eICU critical care databases, a commercial offerings from Truveta, IQVIA, Komodo Health, and Symphony Hea For each dataset, we detail technical implementation approaches, go-to-market strategies, and specific business opportunities ranging from clinical decision sup to real-world evidence generation. Key themes include the convergence of multi data sources, the critical importance of linking public datasets with proprietary clinical data, and the strategic advantages of vertical-specific solutions over hori platforms. We argue that sustainable businesses require not just data access but differentiated domain expertise, proprietary data layering, and strategic go-to-m execution.

The Economics of Healthcare Data and Why Most Companies Get It Wrong

There is something deeply ironic about the state of public health data in Americ have more of it than ever before, much of it genuinely useful and machine-reada sitting in cloud buckets and behind REST APIs waiting to be queried. And yet n health tech companies still treat public datasets like the free breadsticks at Oliv Garden: nice to have around but not something you build a meal around. This is mistake, and potentially a massive one, because there are real businesses to be bu the foundation of public health data if you understand what problems these data actually solve and for whom.

Let me start with the basic economics here because they matter more than most people think. Public datasets have zero marginal cost to access, which sounds gr until you realize that zero marginal cost also means zero barriers to entry for you competitors. Every other smart person in health tech can pull the same data you run the same models, and build similar products. So the question is not whether

can access the data but whether you can do something with it that creates defens value. And that usually means combining public data with something proprietar whether that is unique clinical data from partners, specialized domain expertise lets you interpret the data better than others, or go-to-market relationships that you distribution advantages.

The private data market adds another layer of complexity. Companies like IQVI Symphony Health, Komodo, and the newly ascendant Truveta sell access to clear normalized, and linked patient-level data that can cost anywhere from tens of thousands to millions of dollars annually depending on scale and use case. These datasets offer significant advantages over public data in terms of coverage, granu and timeliness, but they also create a different type of barrier: capital requireme you need to spend six figures just to access the data before you can build anythir better have either venture funding or a clear path to monetization that justifies t upfront investment.

Public Imaging Datasets and the Community Hospital Opportunity

The Google Cloud Healthcare API platform has become something of a develope favorite for imaging-heavy use cases, particularly because it includes datasets lik NIH Chest X-ray collection and the broader Imaging Data Commons repository. are not small datasets either, we are talking about hundreds of thousands of deidentified medical images with varying levels of metadata attached. The techn implementation here is relatively straightforward if you are already in the Googl Cloud ecosystem: you can spin up compute instances, run your ML pipelines dir against the data, and avoid the headache of managing petabytes of imaging data your own infrastructure. The data is generally licensed under Creative Common which means you can use it for commercial purposes with attribution, though yo should always read the specific terms because some subsets have additional restrictions.

The obvious business model here is building diagnostic AI tools that can be vali on public datasets before being deployed in clinical settings. But here is where m companies make their first mistake: they build a general-purpose imaging AI pla that can theoretically work on any radiology image, and then they discover that hospitals do not buy general-purpose tools. Hospitals buy solutions to specific problems. So instead of building yet another "AI-powered radiology platform," y should identify a specific clinical use case where imaging AI can reduce costs or improve outcomes in a measurable way, validate your model on the public datase and then partner with a radiology group or hospital system to get access to their production data for fine-tuning and deployment.

One angle that I think is genuinely underexplored is using these public imaging datasets as a foundation for building clinical decision support tools specifically f community hospitals and rural health systems. These organizations typically do have access to specialist radiologists for every case and often rely on teleradiolog services that are expensive and have variable quality. If you could build a validat model that flags high-priority findings or triages cases for specialist review, you sell that as a service to these facilities at a price point that undercuts traditional teleradiology while providing faster turnaround times. The go-to-market here w likely involve partnerships with hospital purchasing groups or rural health netw rather than trying to sell directly to individual facilities, because the decision-m process at small community hospitals is notoriously slow and relationship-drive

The integration challenge with imaging data is not trivial though. Most hospital systems were built in an era when interoperability meant being able to export a C ROM, and while things have improved with DICOM standards and newer APIs, still end up doing a lot of custom integration work to get your AI model to actua run in the clinical workflow. This is where I think there is an opportunity for infrastructure companies that sit between the AI model developers and the healt systems, essentially providing a normalized API layer that handles the PACS integration, DICOM parsing, HL7 messaging, and all the other plumbing that no wants to build themselves. You could charge both sides for this: model develope

for access to the integration infrastructure, and health systems pay for the orchestration and management layer.

Public Health Surveillance Data and th Government Sales Opportunity

Moving to the CDC Open Data initiative, this is a completely different type of d with different opportunities. The CDC publishes everything from vaccination ra chronic disease prevalence to injury statistics, usually aggregated at the county c state level. This data is incredibly valuable for population health applications, pu health surveillance tools, and for layering additional context onto patient-level c data. The technical access is generally through the Socrata Open Data API, whic pretty developer-friendly if you have worked with REST APIs before. The main g is understanding the aggregation units and how the data has been sampled or adjusted, because the CDC does a lot of statistical modeling and imputation beh the scenes to fill gaps in their surveillance data.

The business opportunity with CDC data is less about building standalone prod and more about using it to enhance other health tech applications. For example, are building a care management platform for chronic disease patients, you could county-level CDC data on disease prevalence and social determinants to help car managers identify which patients are at highest risk based on where they live. O you are building a population health analytics tool for payers or ACOs, you coul layer CDC data onto claims data to show how local health trends correlate with healthcare utilization and costs. The value proposition here is giving your custor additional context that they cannot easily get from their own data alone.

One specific business model that I think could work is building vertical-specific public health dashboards for state and local health departments. Most public he departments are using tools that are a decade or more out of date, and they have limited technical resources to build modern analytics infrastructure themselves. you could create a configurable dashboard platform that ingests CDC data along state-specific data sources and presents it in a way that is actually useful for pub

health decision-making, you could sell that as a SaaS product to state and local h departments. The challenge is that government sales cycles are long and budgets constrained, so you would need to be patient and probably get some initial custo through grants or pilot programs before you could scale it commercially.

The integration challenge with CDC data is mostly about data quality and schen harmonization. Different CDC programs publish data in different formats with different update frequencies and different definitions of what constitutes a repo case or outcome. If you are trying to build a product that combines multiple CD datasets or that combines CDC data with other sources, you end up spending a l time on ETL and data cleaning. This is honestly where most startups building or public data die: they underestimate how much engineering effort goes into maki data usable and they run out of runway before they get to the actual product development. My advice would be to start with a narrow use case that only requi one or two data sources, prove that you can create value with that, and then expa more complex data integration as you grow.

Provider Network Analytics and Insurance Marketplace Data

The HHS HealthData.gov portal is a bit of a mixed bag because it aggregates datasets from dozens of different federal agencies, and the quality and usefulness varies wildly. But buried in there are some genuinely valu datasets, particularly around provider credentials, insurance marketplace data, a healthcare facility characteristics. The [Data.Healthcare.gov] (http://Data.Healthcare.gov) portal specifically has API access for plan and provi information, which is useful if you are building tools for health insurance shopp provider network optimization, or healthcare market analysis. The data tends to more operational and structural rather than clinical, which means the use cases different from what you would do with EHR or claims data.

One business opportunity here that I have seen a few companies pursue with var success is building provider network analytics tools for payers and self-insured

employers. The basic idea is that you combine public data on provider credentia quality metrics, and costs with claims data from the customer to help them opti their provider networks. You might identify high-cost low-quality providers tha should be excluded from narrow networks, or you might find gaps in network coverage where patients are traveling long distances for care. The challenge is th payers and large employers already have analytics teams working on this stuff, sc need to either provide significantly better insights than they can generate intern you need to make the analysis much faster and easier than their current tools all

Another angle is using the marketplace data to build consumer-facing tools for l insurance shopping. The problem with the current state of health insurance sho is that most consumers have no idea how to evaluate plans beyond looking at premiums and deductibles, and they end up making suboptimal choices that cos them money. If you could build a tool that takes into account a consumer's expec healthcare utilization, their preferred providers, and their risk tolerance, you co recommend plans that are genuinely better fits for their needs. The go-to-marke would likely be through partnerships with employers who want to help their employees make better benefits decisions, or through insurance brokers who are looking for tools to differentiate their services. The monetization could be refer fees from plans, subscription fees from employers, or transaction fees from brok

The integration challenge with HHS data is that it sits outside of the clinical wo entirely, so you are not dealing with EHR integration or claims feeds. Instead the challenge is building user interfaces and workflows that make the data accessibl non-technical users. Most of the potential customers for products built on HHS are not going to be data scientists; they are going to be HR managers, benefits consultants, or individual consumers. So you need to invest heavily in product d and user experience, which is a different skill set than most health tech compani have in-house.

Environmental Health Data and the Emerging Multimodal Opportunity

Now we get to SatHealth, which is one of the more interesting new datasets to c out recently because it combines satellite-derived environmental data with socia determinants indicators and claims-based disease prevalence. This is the kind of multimodal dataset that people have been talking about for years but that we ha really seen much of in practice. The initial release covers Ohio and includes thin like air quality measurements, green space availability, food desert indicators, a how those correlate with chronic disease rates. The technical challenge with this is that you need to be comfortable working with geospatial data formats and you to understand how to properly merge environmental exposures with health outc while accounting for confounding variables and temporal lags.

The business opportunity with environmental health data is still emerging, but I there are a few promising directions. One is building tools for public health departments and environmental agencies to identify and address environmental hazards. For example, you could create a system that automatically flags areas wh air quality measurements correlate with elevated asthma hospitalization rates, triggering investigations and interventions. Another direction is clinical risk prediction models that incorporate environmental exposures as features. There i pretty good evidence that environmental factors influence health outcomes independent of traditional clinical risk factors, so adding this data to predictive models could improve accuracy. The challenge is that most healthcare organizat do not currently think about environmental determinants of health in their operational workflows, so you would need to do significant education and chang management to get them to actually use these tools.

The go-to-market strategy for environmental health products is tricky because y selling to organizations that may not have clear budget lines for this type of capa Public health departments are an obvious customer but they tend to have limite funding and slow procurement processes. Health systems are another potential customer, particularly those in communities with significant environmental heal challenges, but you would need to demonstrate clear ROI in terms of reduced hospitalizations or improved population health metrics. A potentially faster patl market might be through partnerships with community health organizations or

health nonprofits that are already working on environmental health issues and tl could use your tools to amplify their impact.

The integration challenge with environmental health data is mostly about geoco and spatial joins. You need to be able to take patient addresses from clinical or c data, geocode them to specific coordinates, and then link them to the relevant environmental measurements. This sounds straightforward but in practice you r into all sorts of issues with address quality, privacy concerns around precise geocoding, and the computational complexity of doing spatial operations at scale also need to think carefully about how to handle temporal aspects, because environmental exposures often have lag effects where the health impact shows u months or years after the exposure.

Clinical Notes and Social Determinants Extraction

The SDOH-NLI dataset represents a newer approach to extracting social determ of health information from unstructured clinical notes using natural language inference. This is important because a lot of the most valuable information abou social determinants is buried in clinical narratives rather than being coded in structured fields. Things like housing instability, food insecurity, and social isol often get mentioned in clinical notes but never make it into the coded problem l structured data fields that most analytics tools rely on. Having a labeled dataset training NLP models to extract this information is genuinely useful for building better social determinants screening and risk stratification tools.

The business model here is probably selling to organizations that are already try do social determinants screening but finding that their current approaches are missing a lot of information. This might be health plans running complex care management programs, ACOs trying to identify high-risk beneficiaries for intervention, or social service organizations trying to connect people with resou The value proposition is that by extracting SDOH information from clinical note can identify needs and risks that would otherwise be invisible. You could build t

a SaaS product that ingests clinical notes through HL7 feeds or FHIR APIs, runs NLP models, and outputs structured SDOH data that can be consumed by other systems.

The integration challenge with clinical NLP is that you need access to clinical n the first place, which means getting business associate agreements in place and handling PHI appropriately. You also need to deal with the fact that clinical wri styles vary dramatically across organizations and even across individual clinician models trained on public datasets may not perform as well on your customer's sp data. This means you probably need some capability for model fine-tuning or adaptation based on customer-specific data, which adds complexity to your prod and deployment process.

Critical Care Research Databases and Academic Validation Opportunities

The MIMIC-IV and eICU databases represent a different category of public data that is particularly valuable for academic research and clinical algorithm validat MIMIC-IV contains deidentified electronic health record data from over 40,000 admissions at Beth Israel Deaconess Medical Center covering the period from 2(2019, while eICU includes data from over 200,000 ICU admissions across 208 hospitals. These datasets are incredibly rich, including everything from vital sig recorded every few minutes to lab results, medications, procedures, and clinical The data is available for free but requires completing a training course on respo use of clinical data and agreeing to a data use agreement.

The business opportunity with critical care databases like MIMIC and eICU is l about building commercial products directly on the data and more about using t for algorithm validation and academic publication before commercializing those algorithms. If you are developing a sepsis prediction model or an ICU mortality score, you can validate it on MIMIC-IV and publish the results in peer-reviewed journals, which gives you credibility when you go to sell the commercial version hospitals. The datasets are also incredibly valuable for training data scientists ar

machine learning engineers who want to work in healthcare, since they provide realistic clinical data without the access restrictions of actual patient data.

One specific business model that leverages these databases is building a consulti services business around helping pharma companies and CROs design better cli trials based on real-world ICU data. You could analyze MIMIC-IV to understanc typical treatment patterns for a specific condition, identify realistic inclusion an exclusion criteria, and estimate likely enrollment rates and dropout rates. Pharn companies pay good money for this type of analysis because it can save them mil of dollars by helping them design trials that are more likely to succeed. The go-t market would be through existing relationships with clinical development teams pharma companies or through partnerships with CROs who need this type of expertise.

The technical challenge with MIMIC-IV and eICU is that the data is messy and requires significant cleaning and preprocessing before you can use it for most applications. Clinical notes need to be parsed, timestamps need to be aligned, ar there are numerous data quality issues that need to be addressed. The good news that there is a large community of researchers who have already worked with the datasets and have published code and documentation on GitHub, so you are not starting from scratch. But you still need people on your team who understand cl data and can navigate the complexities of ICU datasets.

Commercial EHR Data Platforms and th Real-World Evidence Gold Rush

Truveta represents a new model in commercial healthcare data, built as a collect over 30 major US health systems including CommonSpirit, Providence, Trinity Health, and others. As of early 2025, Truveta now includes deidentified EHR dat over 120 million patients, which represents more than one-third of the US popul More importantly, they have just added linked closed claims data for over 200 m patients, creating one of the most comprehensive longitudinal patient datasets available commercially. The data is updated daily, includes clinical notes and me

images, and is structured to meet FDA regulatory standards for real-world evide generation. Recently they also announced the Truveta Genome Project with Regeneron, which aims to sequence the exomes of 10 million volunteers and link genetic data to their EHR records.

The business opportunity with Truveta data is primarily in real-world evidence generation for pharma companies, medical device manufacturers, and payers. Th post-market surveillance studies, comparative effectiveness research, health economics outcomes research, and clinical trial design optimization. The data is expensive to access, probably in the mid to high six figures annually depending o your use case, but it provides a level of completeness and representativeness tha hard to achieve with other data sources. If you are a pharma company trying to understand real-world treatment patterns for a specific indication, or a medical company trying to demonstrate the effectiveness of your device in diverse patien populations, Truveta data is genuinely valuable.

The go-to-market challenge with Truveta is that you need to have relationships w pharma commercial teams, medical affairs, or health economics departments to get in the door. This is not a product you can sell through a website or a sales development rep cold calling. You need someone who understands the pharma business model, who can speak the language of real-world evidence, and who has credibility with the decision-makers. My recommendation would be to partner w established consulting firms or CROs who already have those relationships, rath than trying to build a direct sales motion from scratch.

The technical integration challenge is that Truveta provides data access through own analytics platform rather than giving you raw data exports, which means yo to either do your analysis within their environment or export aggregated results. is actually a feature not a bug from a privacy perspective, but it does mean you c just download the data and use your own tools. They do provide APIs and suppo common analytics workflows, but there is definitely a learning curve.

The Incumbent Commercial Data Vendo and Their Stranglehold on Pharma

IQVIA, Symphony Health, and Komodo Health represent the incumbent comme data vendors that have been selling healthcare data to pharma companies for dec in some cases. IQVIA is the gorilla in the space with their National Sales Perspe and National Prescription Audit datasets that track pharmaceutical sales and prescriptions across retail, institutional, and mail-order channels. Symphony He now owned by ICON, provides similar prescription and medical claims data witl coverage of over 300 million patients. Komodo Health positions itself as having t most comprehensive patient journey data with over 330 million patients and clai capture over 95% of medical claims and over 90% of prescription claims in the U

These vendors are not selling raw data; they are selling curated datasets, analytic platforms, and often consulting services on top of the data. The business model i typically annual subscriptions that can range from hundreds of thousands to mil of dollars depending on the scope of data access and the level of service. The dat include longitudinal patient-level data that has been deidentified and normalize allowing pharma companies to track treatment patterns, understand switching behavior, identify prescriber characteristics, and measure market share.

The opportunity for startups here is not to compete head-to-head with these ven but to build specialized applications on top of their data or to serve markets tha these vendors ignore. For example, you could license data from IQVIA or Symph and build a vertical-specific analytics platform for rare disease drug manufactur who need specialized workflows that the general-purpose platforms do not prov Or you could focus on medical device companies who have been underserved by vendors historically because they focus primarily on prescription drugs. The key finding a niche where the incumbents are not optimizing their product and whe can provide significantly better value to a specific customer segment.

The go-to-market challenge is that these incumbent vendors have deep relations with pharma companies going back years or even decades, and switching costs a high. Your best bet is probably to focus on smaller or mid-size pharma compani

who are not getting good service from the incumbents, or to focus on specific therapeutic areas where you can demonstrate unique expertise. You could also consider a partnership approach where you build applications that sit on top of incumbent platforms rather than trying to replace them entirely.

Synthesis: Building Defensible Businesses on Healthcare Data

Across all of these opportunities, there are some common themes that show up repeatedly. First, the highest value usually comes from combining public dataset with proprietary data sources rather than trying to build products on public dat alone. Public data gives you a foundation and lets you demonstrate initial value, create truly differentiated products you need to layer in data that your competit not have. Second, successful businesses built on healthcare data tend to be highl verticalized rather than horizontal platforms. The companies that win are the or solving specific problems for specific customer segments rather than building ge analytics tools that could theoretically be used by anyone. Third, go-to-market is harder than product development when you are working with healthcare data. T data access is tractable, the modeling is tractable, but actually finding customers will pay for what you have built and integrating into their workflows is where m companies struggle.

One final thought on defensibility, because this is the question that investors alv ask about businesses built on healthcare data. If everyone has access to the same public datasets, or if multiple vendors are selling similar commercial data, what competitors from copying your product? The answer is that your moat comes fro everything that is not the data itself: your domain expertise that lets you identify right problems to solve, your product and user experience design that makes the accessible to non-technical users, your go-to-market relationships and partnersh that give you distribution advantages, your proprietary algorithms or data source you layer on top of the base datasets, and the operational workflows and integrat you build that make your product sticky once customers start using it. Healthca

data is a starting point, not a destination, and the companies that understand tha
the ones that will build sustainable businesses in this space.

The reality is that most healthcare data opportunities require either significant
upfront capital to license commercial datasets, or significant technical expertise
properly clean and utilize public datasets, or both. The companies that succeed a
the ones that can either raise enough venture capital to fund the upfront investm
or that can find creative ways to bootstrap by starting with public data and
progressively adding proprietary capabilities as they generate revenue. The ange
investing opportunity here is identifying teams that have the right combination
domain expertise, technical capability, and go-to-market savvy to actually execut
these opportunities rather than just having an idea about what could be built wi
data.

The opportunity in healthcare datasets is real but it requires a different approac
than most venture-backed health tech companies are taking. You need to start w
specific problem and customer segment, use available data sources to build an in
solution, layer in proprietary capabilities to create differentiation, and focus
relentlessly on go-to-market execution. If you can do all of that, there is genuine
to be created and captured. But if you think you can just download some public
datasets or license some commercial data, train some models, and sell generic
analytics software, you are going to have a hard time. The data might be accessib
but the opportunities are not obvious, and that is exactly why they are worth pur

---

If you are interested in joining my generalist healthcare angel syndicate, reach o
treyrawles@gmail.com or send me a DM. We don't take a carry and defer annual
for six months so investors can decide if they see value before joining officially.
Accredited investors only.

1 Like

← Previous

Nex

Discussion about this post

Comments Restacks

Write a comment...