Price Transparency as Infrastructure: Building Defensible Businesses on CA Data

NOV 17, 2025 • PAID

DISCLAIMER: The views and opinions expressed in this essay are solely my own and reflect the views, opinions, or positions of my employer, Datavant, or any of its affiliat

If you are interested in joining my generalist healthcare angel syndicate, reach o treyrawles@gmail.com or send me a DM. Accredited investors only.

TABLE OF CONTENTS

ABSTRACT

The Consolidated Appropriations Act of 2021 created a price transparency mand that most people still don't understand. Employers and health plans must now p machine-readable files showing negotiated rates with every provider in their net Three years into implementation, we have one of the most comprehensive health pricing datasets ever assembled, and almost nobody is using it effectively. This e examines what data actually exists, how to access it, which business models show traction, and why enforcement inconsistency creates both opportunity and risk. health tech investors, this represents a rare moment where regulatory infrastruc has been built but commercial applications remain nascent. The window for firs mover advantage is open but closing as larger players begin to recognize the asse they're sitting on.

Introduction: The Accidental Infrastructure Play

So here's the thing about the CAA price transparency requirements that nobody talks about: they weren't designed to create a new data infrastructure layer for healthcare. They were designed to shame health plans and employers into compe on price by exposing the absurd variation in what they pay for identical services. theory was pretty straightforward - if you force plans to publish what they're act paying every provider for every service, market forces would kick in and prices w normalize. Employers would look at the data and realize they're getting ripped o Patients would shop around. Providers charging 10x what their competitor charg for the same MRI would have to justify it or lose business.

That's not really what happened. What actually happened is we accidentally buil of the most comprehensive datasets of healthcare pricing in American history, an almost nobody knows it exists or how to use it. Which, if you're reading this as a health tech investor, should make your ears perk up because that's basically the definition of an opportunity.

The Consolidated Appropriations Act got passed in December 2020 as one of th massive omnibus bills that everyone votes for because it funds the government a has COVID relief money in it. Buried in there were these transparency provisior came out of the hospital price transparency rules from 2019. The hospital rules s hospitals have to publish their chargemasters and negotiated rates. The CAA rul said the same thing but for health plans and self-insured employers. The enforce date was July 1, 2022 for the plan-level transparency requirements.

Now we're in November 2025 and the data landscape is starting to mature. Not because enforcement is great - it's actually pretty inconsistent - but because eno plans have published enough files that you can start building real products on tc it. The question isn't whether the data exists anymore. The question is what you with it.

What Actually Got Built: Understanding the Data Landscape

Let me walk you through what actually has to get published because this is wher most people get confused. There are three separate disclosure requirements in t CAA and they're all different.

First, there's the machine-readable files showing in-network negotiated rates. Tl the big one. Every health plan has to publish a file that shows every negotiated r with every provider in their network for every service code. This means you can that Aetna pays Hospital A $1,200 for a knee MRI but pays Hospital B $3,800 fo exact same CPT code in the same market. The files are supposed to be in JSON f following a specific schema that CMS published. They're supposed to be update monthly. They're supposed to include every single negotiated rate for every cove item and service.

Second, there's historical out-of-network allowed amounts. Plans have to publis they've actually paid out-of-network providers over the past year for different se codes. This is supposed to be aggregated data showing the range of amounts pai

it gives you a window into what plans are willing to pay when they don't have a negotiated contract.

Third, there's the cost-sharing disclosure requirement. This one is less about bu data and more about patient-facing tools. Plans have to provide members with personalized cost estimates for any covered service through either an internet-b tool or in paper form on request. This creates the requirement for plans to build estimator tools that can actually calculate out-of-pocket costs based on the patie specific benefit design and where they are in their deductible.

The in-network rate files are where the real action is for investors. These files ar massive - we're talking about multi-gigabyte JSON files in many cases. A large national plan might have dozens or hundreds of these files covering different products, different regions, different plan years. Each file contains millions of rc pricing data. UnitedHealthcare's rate files reportedly total over 200 terabytes wh you add up all their different products and regions. That's not a typo.

The data structure follows a schema that CMS published but the implementatio varies wildly. Some plans publish clean, well-structured files that are relatively e parse. Others publish files that are technically compliant but practically unusab without serious data engineering work. There are plans that split their data into thousands of small files. There are plans that publish files so large they're hard t download. There are plans that use provider identifiers that don't match up to a standard taxonomy. It's a mess.

But here's the interesting part - the mess itself creates a moat. If you can build t infrastructure to reliably ingest, normalize, and query this data at scale, you hav something valuable. The barrier isn't that the data is secret anymore. The barrie that the data is so fragmented and inconsistent that it takes real engineering effc make it useful.

Let me give you a concrete example of what's in these files. Say you want to knov different plans pay for a colonoscopy in Atlanta. You'd need to download the in- network rate files for every major plan operating in Georgia. You'd need to parse

the rates for CPT code 45378 (colonoscopy with biopsy) for every provider. You'd to match the provider identifiers to actual facility names and addresses. You'd n account for the fact that the same provider might have different rates for differe plan products from the same carrier. You'd need to handle cases where rates are expressed as percentages of Medicare rather than dollar amounts. You'd need to with billing codes that have modifiers or special rules.

Now multiply that across every procedure code and every market and every plan That's the dataset we're talking about. And that's before you get into any of the interesting analytical work like identifying pricing patterns or outlier providers building recommendation engines.

The other piece of this that matters is the allowed amount data for out-of-netwo claims. This data is generally less granular - it's supposed to show the distributi amounts paid rather than claim-level detail. But it gives you a benchmark for wh plans actually pay when they don't have leverage. The spread between in-networ negotiated rates and out-of-network allowed amounts tells you something about value of network contracts and where plans have pricing power versus where the just accepting whatever providers charge.

Getting Your Hands on the Data: A Practical Guide

Okay so how do you actually access this data if you want to build something on i There are basically three approaches and they all have tradeoffs.

The first approach is to go directly to the source. Health plans are required to m these files publicly accessible without requiring any authentication. Most plans the files on a dedicated transparency portal. You can literally just go to a plan's website, navigate to their price transparency page, and start downloading files. UnitedHealthcare has a transparency portal. Anthem has one. Cigna has one. Ev major plan has to have one.

The catch is that finding these portals is not always straightforward and navigat them is often deliberately difficult. Plans aren't exactly incentivized to make thi easy to access. They're complying with the letter of the law but not the spirit. So plans bury their transparency pages several clicks deep in their site navigation. S plans use file naming conventions that make it hard to figure out which file cont the data you want. Some plans rate-limit downloads or make you solve CAPTCH Some plans split data into so many separate files that bulk downloading become technical challenge.

There's also the small problem that if you want comprehensive data across multi plans, you're talking about petabytes of files. The download alone would take we or months depending on your bandwidth. Then you need somewhere to store it a the compute infrastructure to process it. This is not a laptop-scale problem.

The second approach is to use a data aggregator that's already done the heavy lif There are a handful of companies that have built infrastructure to systematically download, normalize, and index the transparency data from major plans. These companies essentially provide a cleaned-up version of the raw data through APIs data extracts. Turquoise Health is probably the most well-known player here. Healthcare Bluebook has integrated transparency data into their platform. Ribb Health has transparency data as part of their provider data offering. There are ot emerging.

The tradeoff with aggregators is cost and coverage. These companies are providi valuable service so they charge for it. Pricing varies but you're generally looking five or six figures annually for API access depending on your use case and volum Coverage also varies - no aggregator has 100 percent of plans because the long ta regional and small plans is massive. But the major aggregators will cover the big national plans and most significant regional plans which gets you probably 70-8(percent of covered lives in most markets.

The third approach is to use one of the free or low-cost research datasets that academics or nonprofits have assembled. HCCI (Health Care Cost Institute) has started incorporating transparency data into their research datasets. Some acade

research groups have built scrapers and published data extracts. The tradeoff he that these datasets are usually less comprehensive and less frequently updated tl commercial offerings, but they're accessible if you're doing exploratory work or building a prototype before you invest in commercial data access.

For investors evaluating a startup that's building on transparency data, the key questions are: What's your data strategy? Are you building your own ingestion pipeline or relying on a vendor? If you're building your own pipeline, do you hav engineering resources to maintain it as schemas change and plans update their f If you're using a vendor, what happens to your unit economics as you scale? How defensible is your data moat if anyone can access the same underlying data?

My take is that most startups should start with an aggregator to validate the bus model and then evaluate whether to bring data ingestion in-house once you have revenue and scale. Building your own pipeline is a significant engineering lift th distracts from building the actual product unless data engineering is your core competency. But there's a point at which vendor fees become prohibitive and hav more control over data quality and freshness becomes strategically valuable. Tha inflection point is probably somewhere around a few million in ARR depending how data-intensive your product is.

The Business Model Playbook: What Works and What Doesn't

Let's talk about what you can actually build with this data. I've seen or heard pit for probably a dozen different business models at this point and there's starting some pattern recognition about what has legs versus what sounds good in a pitcl doesn't really work in practice.

The most obvious use case is price comparison and shopping tools for patients. T theory is simple - you show patients that Provider A charges $1,500 for a service Provider B charges $500 for the same thing, the patient chooses the cheaper opti everyone saves money. The problem is that this has been tried for years in health with limited success because patients don't actually shop for healthcare services

same way they shop for TVs or airline tickets. They want convenience, they trust doctor's referral, they're often in pain or scared, and the out-of-pocket cost diffe after insurance might not be that dramatic even if the negotiated rate difference huge.

That said, there are contexts where price shopping actually does work. Elective procedures where the patient has time to shop and the service is commoditized. Imaging and lab work where the experience is basically the same regardless of w you go. High-deductible situations where the patient is paying the full negotiate out of pocket so the cost difference actually matters to them. But you need to be realistic about what percentage of healthcare spend fits that profile. It's not noth but it's not everything either.

The more interesting B2C play in my view is using transparency data to power advocacy and navigation services. If you can combine transparency data with cla data and network data and clinical appropriateness criteria, you can build tools proactively identify opportunities for patients to save money without sacrificing quality. Instead of making the patient do the work to search and compare, you d work for them and push recommendations. This is what companies like Accolad Castlight and Quantum Health do to some degree, though most of them were bu before transparency data existed so they're retrofitting it into existing platforms

On the B2B side, the opportunities are stronger. Employers and benefit consulta are desperate for tools to understand whether they're overpaying and which pro are outliers. The transparency data finally gives them the benchmarks to have informed conversations with their health plans and TPAs about network design steerage programs. If you can show a benefits team that they're paying 2x the ma rate for orthopedic surgery at a particular hospital system, that's actionable intelligence they didn't have before.

There's a whole category of B2B analytics tools emerging here. Tools that ingest transparency data and combine it with an employer's own claims data to identify opportunities for network redesign or steerage or reference-based pricing. Tools benchmark an employer's plan performance against peer plans or market averag

Tools that identify providers with pricing or quality outliers and help employers decide whether to keep them in-network or not.

The challenge with B2B analytics is that you're selling to a buyer (benefits teams is traditionally not very sophisticated about data and analytics, has limited budg and is already overwhelmed with vendor pitches. The sales cycle is long and the contract sizes are often smaller than you'd like unless you're selling to Fortune 5 companies. But the need is real and the transparency data makes the pitch a lot compelling than it was when you were just showing them their own claims data dashboard.

Another B2B opportunity is selling directly to health plans themselves. This sou counterintuitive - why would plans pay for tools built on data they're required t publish? But plans need help making sense of their own transparency data. They tools to audit their own files for compliance. They need analytics to understand their negotiated rates compare to competitors. They need patient-facing cost estimator tools to comply with the member transparency requirements. There ar several startups building compliance and analytics tools specifically for plans.

Provider-facing tools are another category. Providers don't have good visibility i what different plans are paying them relative to their competitors. The transpare data creates an opportunity for benchmarking and rate negotiation tools aimed a providers. Show a provider that they're being paid 30 percent below market for a particular service, that's useful information when they're negotiating their next contract with the plan. This is basically a RevCycle analytics play with transpare data as the special sauce.

The business model that I think has the most potential but is also the hardest to execute is using transparency data as a foundation for new payment models or network products. Imagine a direct contracting entity that uses transparency da identify high-quality low-cost providers and build narrow networks around them a reference-based pricing product that uses transparency data to set payment benchmarks and then helps employers manage the operational complexity of RB a bundled payment intermediary that uses transparency data to identify which

providers can deliver episodes of care at attractive prices and then contracts dire with those providers on behalf of employer groups.

These models are capital-intensive and operationally complex. You're not just building a data product, you're building what amounts to a health plan or a payr intermediary or a network. But the transparency data creates new possibilities because it reduces information asymmetry. You can build network products witl having to go through years of negotiating rates with providers because you alrea know what they're accepting from other plans. You can price bundled payment products more accurately because you know what the component services actual cost in-network.

The Enforcement Problem and Why It Matters

Here's the uncomfortable truth about CAA transparency: enforcement is wildly inconsistent and lots of plans are not compliant. CMS is the enforcement agency they're supposed to be auditing plans for compliance and issuing penalties for violations. The penalties can be significant - up to $100 per day per violation per affected individual, which could theoretically add up to millions of dollars for a plan that's not compliant.

In practice, CMS has been pretty gentle. They've mostly taken an educational approach - sending warning letters, giving plans time to come into compliance, treating this as a learning process rather than a strict enforcement regime. There been some penalty assessments but they're relatively small and infrequent given scale of non-compliance. A lot of plans are publishing files that are technically present but don't actually meet the requirements. Files that are missing significa portions of data. Files that are structured in ways that make the data unusable. I that aren't updated monthly like they're supposed to be.

This creates a few problems for anyone trying to build a business on transparenc data. First, data completeness is inconsistent. You can't assume that a plan's published files contain all their negotiated rates. There might be whole provider

groups or facility types that are missing. This means you need to be careful abou making definitive claims based on the data - you're analyzing what's published, v might not be what actually exists.

Second, data quality varies tremendously. Some plans clearly have mature data governance processes and publish clean files with good metadata and consistent identifiers. Other plans are dumping raw extracts from their claims systems witl much cleanup. If you're building a product that depends on the data being accu and complete, you need to invest heavily in data quality checks and normalizatio

Third, there's political risk. The transparency requirements came out of a Repul administration but have been maintained and expanded under a Democratic administration, which suggests they have bipartisan support. But there's always that enforcement priorities could change or that plans could successfully lobby t weaken the requirements. If you're building a company that's entirely dependen transparency data continuing to be available and improving in quality, that's a ri factor investors need to consider.

The flip side of weak enforcement is that it creates opportunity. Plans are slowly getting better at compliance because the requirements aren't going away and the reputational risk of being publicly non-compliant is increasing. There are consultancies now that specialize in helping plans achieve transparency complia The data is getting better over time. If you can build a business now while the da still messy and incomplete, you'll have a significant advantage when the data qua improves because you'll have already solved all the hard engineering problems.

Future State: Where This Goes Next

Let's talk about where this is headed because the current state of transparency d very much a transitional phase and the end state is likely to look quite different what we have now.

In the near term, I expect continued improvement in data quality and completen plans get better at compliance and CMS gets more serious about enforcement. T

Trump administration is back in power as of January 2025 and while their specif priorities on transparency aren't entirely clear yet, the transparency push origin came from the first Trump administration so there's reason to think enforcemen might actually get stronger rather than weaker. But that's speculative.

More immediately, I expect we'll see consolidation in the data aggregator space. now there are maybe half a dozen companies trying to be the definitive source fo transparency data and there's probably only room for two or three to be viable lo term. The ones that survive will be the ones that can demonstrate the best data quality, the most comprehensive coverage, and the most useful enrichment on to the raw data. There will be M&A activity here - wouldn't surprise me to see one big healthcare data incumbents like IQVIA or Komodo or Symphony acquire a transparency data aggregator to add it to their portfolio.

On the product side, I expect we'll see transparency data get baked into existing workflows rather than being a standalone product category. Cost estimator tools patient portals. Network adequacy analytics in TPA platforms. Rate benchmarki provider revenue cycle systems. The data becomes table stakes infrastructure rat than a differentiated product. That's good for adoption but challenging for start trying to build standalone transparency businesses.

The really interesting question is whether transparency data enables genuinely r market structures. Does it make it easier for new entrants to build health plans networks because they don't have to negotiate every rate from scratch? Does it c opportunities for bundled payment models or reference-based pricing to scale b their current niche? Does it put downward pressure on healthcare prices becaus employers finally have the tools to effectively steward their spending?

My instinct is that transparency alone isn't sufficient to transform market struct but it's necessary. You still need to solve all the hard operational problems arou claims processing and provider contracting and member experience. But transpa data removes one of the key information barriers that has historically made heal markets so dysfunctional. Combined with other trends - the shift to self-insuran the growth of high-deductible plans, employers getting more sophisticated abou

benefits strategy - transparency data could be part of a larger shift toward more functional price competition in healthcare.

Conclusion: The Window Is Closing

If you're looking at the healthcare investment landscape and trying to identify a where there's greenfield opportunity with regulatory tailwinds and limited competition, transparency data is one of the most interesting spaces right now. T infrastructure has been mandated by regulation. The data exists and is improvin commercial adoption is still early and there aren't clear winners yet in most of th business model categories I described.

That said, the window for early-stage investment is probably only open for anoth year or two. The big data companies are starting to pay attention. The health pla incumbents are starting to build their own tools. The successful early startups a starting to raise growth rounds and capture market share. If you're going to inve this space, now is the time.

The companies I'd be most excited about are ones that combine transparency da with other proprietary data assets to create something defensible. Transparency plus claims data plus clinical data plus network data plus member engagement - a much stronger moat than transparency data alone. I'd also be looking for comp with go-to-market strategies aimed at sophisticated buyers who have budget and authority to make purchasing decisions - large self-insured employers, benefit consultants, health plan innovation teams, provider CFOs.

I'd be cautious about pure-play consumer shopping tools because the behavior c is hard and the incumbents have distribution advantages. I'd be cautious about anything that's entirely dependent on continued aggressive enforcement because that's not a sure thing. I'd be cautious about anything with long sales cycles and contract sizes unless the unit economics are really compelling.

But fundamentally, this is the rare situation in healthcare where regulation has created a new data commons and the commercial applications are lagging behin

data availability. That's an opportunity. The entrepreneurs who can figure out hc turn transparency data into products that meaningfully change decision-making employers, for patients, for providers, for plans - those are the ones who will bui valuable companies. And for investors, the time to get in is before everyone else realizes what just got built.

← Previous

Nex

Discussion about this post

Comments Restacks

Write a comment...