Why Healthcare Public-Data Startups Win on the Adjudicated Entity Graph and Not the Thousand Scrapers: Ownership, Fraud, Rollup, and Distres Intelligence for Hospice, Behavioral Health & Post-Acute

JUN 26, 2026 • PAID

🎧 Podcast episode for paid subscriber only. Also available on Spotify.

Thoughts on Healthcare Markets & Technology

Why Healthcare Public-Data Startups Win on the Adjudicated Entity Graph & Not the Thousand Scrapers: Ownership, Fraud, Rollup, and Distress Intelligence for Hospice, Behavioral Health, and Post-Acute

The ‘we aggregate a thousand public sources’ pitch is dead. AI coding tools just made ingestion cheap. Here’s what that means for every healthcare data startup building right now…

Listen now

a day ago · Thoughts on Healthcare

To listen to paid episodes in Apple or Spotify, link your Substack subscription via the s settings on those platforms (instructions inside the Substack app under Subscriptions →

Podcast).

Abstract

Cheap code generation has quietly nuked the part of healthcare data work that u to feel hard. Writing source-specific crawlers, reverse-engineering undocumente schemas, parsing ugly PDF tables, and babysitting connectors when a state webs redesigns itself is now mostly a tooling problem, not a team problem. So the old "we aggregate a thousand public sources," is no longer a moat. Anyone can aggre thousand sources by next Tuesday. What stays expensive, and therefore defensib the adjudicated temporal entity graph sitting underneath: durable internal IDs f people, organizations, facilities, enrollments, locations, legal entities, owners, ar products, all stitched across NPIs, CCNs, TINs, CLIA numbers, and state licens with valid-from and valid-to dates, source-document provenance, confidence scc and a growing pile of human corrections. This essay walks through the dozen pu data product ideas worth building (ownership graphs, fraud and integrity intel, radar, distress prediction, ghost-network truth, site-of-care migration, trial-site feasibility, device early warning, regulatory catalysts, workforce capacity, post-ac counterparty scoring, and environmental demand), ranks them by buyer urgency versus legal blast radius, and argues for one specific opening move. Pick a single opaque, fragmented sector such as hospice, behavioral health, home health, or sl nursing. Answer four narrow questions about ownership, affiliation, risk, and momentum. Sell the workflow, not the dashboard. Let one buyer's revenue fund graph that every later buyer rents.

Table of Contents

One. The bet hiding inside “we aggregate a thousand public sources”

There is a whole category of healthcare analytics companies whose entire reason existing is that nobody can answer a simple-sounding question without doing tw weeks of forensic accounting. Who actually owns this hospice. Is this nursing ho about to fail. Has the guy applying for enrollment ever been attached to an exclu provider under a slightly different name. None of these has a single dataset that tells you the answer, which is exactly why somebody will pay real money for the answer. The good opportunities all rhyme. The buyer is making a high-dollar de wrapped in risk, revenue, an acquisition, a contract, or regulatory exposure. No public file closes the loop. The underlying entities are a nightmare to reconcile because the same real-world thing shows up as a dozen NPIs, three DBAs, two p legal names, a management company, and a real-estate LLC that exists only to ho building. And the useful output is a narrow operational answer, not yet another general-purpose healthcare data lake nobody asked for.

The trap, and it is a comfortable trap, is to believe the work is the aggregation. I like work. It produces impressive slide counts. CMS alone will keep a team busy quarter just normalizing what it already publishes, and the published hospital ownership file runs to roughly one hundred fifty thousand rows before anyone to a county clerk's office. But aggregation is the part that just got cheap, and cheap things do not protect margins. The defensible asset is what happens after the raw records land: the resolved identities, the longitudinal graph with dates on every the derived scores, and the accumulated record of every time a human looked at records and decided whether they were the same thing. That correction history i part a competitor cannot clone by pointing the same crawler at the same website

← Previous

Nex