The Data Stack That Catches Crooks: Linking Open Datasets to the New Medicaid Spend Data, Why Home Health Is a Fraud Paradise, and How t Build a Business on Top of All of It

FEB 15, 2026

Share

Table of Contents

Abstract

- The new HHS Medicaid provider spending dataset (NPI x HCPCS x month, 201 2024, 227M rows) is powerful alone and exponentially more useful when joined against other free public datasets

- Key linkable datasets: NPPES provider registry (8.6M+ providers, entity forma dates, authorized officials), OIG exclusion list (sanctioned providers and individ CMS Open Payments (Sunshine Act financial relationships), PECOS Medicare enrollment data, SAM.gov federal debarment records, state corporate registry filings, Census TIGER geographic data, HUD housing data, B employment statistics

- Home health is the highest-spend taxonomy in the dataset at $288B+ and the h fraud-density category per dollar due to: zero verifiable clinical artifacts, self-

attesting documentation, caregiver identity impossible to audit at scale, and fede matching rate economics that give states weak policing incentives

- The entrepreneur opportunity is a hybrid model: open source data analytics as top-of-funnel fraud signal generator, wrapped with human investigative services tam legal partnerships, and outcome-based government contracts

- False Claims Act whistleblower recoveries returned $2.80 per $1 spent on enforcement in the HCFAC program; the qui tam relator share is 15-30% of reco funds

- Business model options range from SaaS sold to MCOs and state Medicaid age to investigative services on contingency, to a qui tam legal referral engine, to a fu integrated fraud-to-recovery platform

The Datasets Worth Linking and Why

The Medicaid provider spending dataset that dropped last week is genuinely nov but it is also genuinely incomplete as a standalone fraud detection tool. What it you is a time series of how much money flowed from Medicaid to a specific provi NPI for a specific HCPCS procedure code in each month from January 2018 thro December 2024. What it does not give you is almost everything else you need to whether that billing pattern represents fraud, waste, or legitimate healthcare de The fraud detection signal lives in the gap between what the spending data show what a comprehensive picture of that provider's real-world existence looks like. Closing that gap requires joining the Medicaid data against a stack of other pub datasets that are all free, all downloadable, and all dramatically underutilized by anyone outside the payment integrity industry.

The most important linkable dataset by a wide margin is the NPPES provider re maintained by CMS and publicly downloadable in full from [download.cms.gov] (http://download.cms.gov). NPPES contains registration records for every provid with a National Provider Identifier in the United States, which as of 2025 is over million records. The fields that matter most for fraud detection are not the obvi

ones. The NPI itself is useful as a join key. The authorized official name and con
information for organizational providers is more useful, because it lets you iden
whether the same individual appears as the authorized official across multiple N
registrations, a pattern that shows up consistently in bust-out schemes where
operators open successive LLCs under their own name or under family members
entity formation date is arguably the single most valuable field in the whole regi
for fraud purposes, because the new-entity-plus-rapid-billing-escalation pattern
reliably predictive of fraud in the behavioral health and home care taxonomies th
functions almost as a rule rather than a signal. An LLC that did not exist eightee
months ago and is now billing Medicaid at the 95th percentile for its taxonomy i
state has a prior probability of fraud that dwarfs almost any other indicator.

The OIG exclusion list is the second essential join. The Department of Health an
Human Services Office of Inspector General maintains a publicly downloadable
individuals and entities that have been excluded from participation in federal
healthcare programs, typically as a result of fraud convictions, license revocation
other misconduct. The list is searchable and downloadable at [oig.hhs.gov]
(http://oig.hhs.gov) and is updated monthly. The most common fraud pattern it e
detection of is NPI laundering: an individual who was personally excluded from
Medicaid billing opens a new organizational entity under a spouse, parent, or bu
partner's name, obtains a new organizational NPI, and resumes billing under the
entity. The exclusion list alone cannot catch this pattern because it tracks indivi
by name and Social Security number, not by organizational affiliation, but when
joined against NPPES authorized official data it becomes significantly more pow
An entity whose authorized official shares a surname and address with an exclud
individual is not proof of fraud but it is a screening flag worth following up on.

CMS Open Payments, the dataset created by the Physician Payments Sunshine A
tracks financial relationships between pharmaceutical manufacturers, medical d
companies, and healthcare providers. At first glance this seems unrelated to Mec
home health fraud. In practice it is useful for a specific subset of fraud patterns
involving referral schemes, particularly in durable medical equipment, specialty
pharmacy, and behavioral health service lines where manufacturers pay provider

referrals or consulting arrangements that serve as kickback vehicles. A home he
agency operator who appears as a recipient of significant Open Payments transfe
from a DME supplier while simultaneously billing Medicaid for home health ser
at anomalous rates has a fact pattern worth examining even if neither data point
conclusive alone.

PECOS, the Medicare Provider Enrollment Chain and Ownership System, is CM
enrollment database for Medicare participation and is publicly available through
data.cms.gov. It contains provider enrollment dates, practi
location histories, specialty classifications, and reassignment of benefits relation
that show which individual practitioners have routed their billing through whic
organizational entities. The reassignment data is particularly valuable for detect
the organizational shell game that sophisticated fraud operators play. A physical
therapist who reassigns their Medicare billing through six different LLCs over a
year period, each of which also happens to have been a high Medicaid biller duri
existence, is a pattern that PECOS makes visible in a way that the Medicaid sper
data alone cannot.

SAM.gov, the federal System for Award Management, maintain
database of entities that have been debarred or suspended from federal contract
and program participation. It overlaps partially with the OIG exclusion list but c
a broader range of federal programs and includes some individuals and entities t
appear on one list but not the other. For Medicaid fraud purposes it is most usef
a supplementary screen rather than a primary signal, but the crosswalk between
exclusions and active Medicaid billers has historically surfaced cases that the OI
missed.

State corporate registry data is technically fifty-one separate datasets rather thar
but the major states all publish searchable corporate registration records that id
when an LLC was formed, who its registered agent and members are, and wheth
currently in good standing. Several states have made this data downloadable in b
For the states that haven't, it is often scrapeable through public records request
commercial data vendors who aggregate it. The corporate registry data is the pie
the puzzle that closes the entity relationship graph. When you know from NPPE

a Medicaid billing entity was formed in 2022 and when you know from the state corporate registry that its registered agent is the same person who was the regist agent for three other LLCs that billed Medicaid heavily between 2018 and 2021 then dissolved, you have a fact pattern that is not just anomalous but narratively coherent as a fraud scheme.

Census TIGER geographic data and HUD housing datasets round out the stack i less obvious but analytically important way. A provider claiming to deliver home health services at a residential address that appears in HUD data as a vacant lot, commercial property, or a federally subsidized housing unit with no registered healthcare operations is a geographic implausibility check that costs nothing to and surfaces a meaningful share of phantom billing fraud. The van in rural New Mexico billing 1,006 claims per workday would have failed a basic geographic plausibility screen if the claimed service delivery addresses had been cross-refer against population density and dwelling unit data for that provider's service area

Home Health: A Perfect Fraud Ecosyste

Understanding why home health dominates both the legitimate Medicaid spend taxonomy and the fraud pattern taxonomy requires understanding what home he actually is as a service category and what the structural incentives look like from angle. The short answer is that home health is a perfect fraud ecosystem not bec fraudsters are unusually clever but because the program was designed in a way t makes fraud the path of least resistance and legitimate oversight nearly impossib scale.

The core problem is that home health services, including personal care, attendar care, and home-based behavioral health support, are fundamentally unverifiable any of the mechanisms that work for other healthcare claim types. When a physi bills Medicaid for a surgery, there is an operative note, an anesthesia record, a fa record, a post-operative nursing note, and typically imaging or pathology results collectively make it nearly impossible to bill for a surgery that did not happen. W a pharmacy bills Medicaid for a prescription, there is a dispensing record, a

prescriber NPI, a patient signature or delivery confirmation, and a drug supply c
that leaves multiple independent verification points. When a home health aide b
Medicaid for four hours of personal care services delivered to a beneficiary in th
home, there is a caregiver attestation, a supervisory visit note that in most states
required only quarterly, and perhaps an Electronic Visit Verification timestamp
state has implemented EVV well. That's it. The service itself is a human interact
a private residence that leaves no independent evidence of having occurred.

Electronic Visit Verification, mandated by the 21st Century Cures Act and requ
all states by 2020 for personal care services and by 2023 for home health services
supposed to address exactly this problem. EVV systems require caregivers to che
and out of visits electronically, typically via a smartphone app or telephonic syst
creating a timestamp and often a GPS location record. The implementation reali
been deeply uneven. States had significant flexibility in how they implemented E
which vendors they chose, and how rigorously they enforced compliance. In som
states EVV data is submitted to managed care organizations who are supposed t
match it against claims before payment. In practice the matching is often done
retrospectively after payment has already been made, which converts EVV from
payment fraud prevention tool into a post-payment audit trigger. In other states
compliance is treated as a documentation requirement that generates a correctiv
action plan when violated rather than a payment denial, which means a provider
never submits EVV data faces a compliance letter rather than a clawback. The D
Medicaid spending dataset does not include EVV data, which means analysts wo
from the public data cannot directly assess whether billed visits have EVV
confirmation. That gap is significant.

The caregiver workforce dynamics of home health create additional fraud
vulnerability that is structural rather than incidental. The industry is characteri
high turnover, low wages, part-time employment, and significant use of gig-style
arrangements. A home health agency operator who wants to commit billing frau
two basic choices: bill for services that were partially delivered at a higher unit c
than actually occurred, or bill for services that were never delivered at all using
caregiver names and real beneficiary names without any actual service taking pla

Both schemes are enabled by the labor structure of the industry. In the partial delivery scheme, caregivers who show up for two hours get billed as four, with tl caregiver sometimes complicit and sometimes unaware that their attestation is b inflated by the agency billing department. In the phantom billing scheme, the operator may employ a handful of actual caregivers to create a legitimate-lookin operation while billing far beyond what that workforce could physically deliver. physical impossibility threshold that flags the van in New Mexico at 1,006 claim workday is the extreme version of this pattern. The more common version is an a with twelve W-2 employees billing for a volume of visits that would require thir workers, which looks unusual in a staffing audit but does not surface in billing d alone without workforce size as a denominator.

The Medicaid managed care structure, which now accounts for roughly two-thir total Medicaid spending nationally, creates a diffusion of accountability that hor health fraud operators have exploited systematically. In a fee-for-service Medica world, the state Medicaid agency pays the provider directly and receives the clai directly, creating at least the theoretical possibility of state-level anomaly detect In managed care, the state pays a per-member-per-month capitation to the MCC MCO pays the provider, and the MCO is nominally responsible for fraud detecti within its network. In practice MCO fraud detection programs vary enormously sophistication and most are focused on the highest-dollar, most obvious patterns because the economics of payment integrity under capitation are different from for-service. Under capitation the MCO absorbs the cost of fraud from its medica ratio unless it can recover from the provider, which creates incentive to detect ar claw back fraud but also creates incentive to keep enrollment high and networks broad in ways that can compete with rigorous credentialing. The result is a syste where the entity theoretically responsible for catching home health fraud is also entity that signed the contract with the home health agency and has commercial relationships with its operators.

Federal matching rate economics complete the fraud-enabling picture. Because t federal government pays between fifty and ninety cents of every Medicaid dollar depending on the state's Federal Medical Assistance Percentage, states bear only

fraction of the cost of fraudulent payments made within their programs. A state a 70-30 federal-state matching ratio that allows $100M in fraudulent home healt billing to persist only loses $30M from its own budget. The federal taxpayer abs the rest. This creates a documented pattern where states with higher federal ma percentages have historically had weaker fraud detection infrastructure, not bec state officials are corrupt but because the incentive structure makes aggressive f enforcement a worse financial proposition than it appears from the outside. Fixi this requires either changing the matching rate structure in ways that give state stronger financial stake in their own program integrity, or creating federal detec infrastructure that operates independently of state incentives. The new public d is a meaningful step toward the latter.

The Entrepreneur's Playbook for Monetizing Open Source Fraud Detecti

The business opportunity created by the Medicaid spending dataset combined w the other public datasets described above is real, meaningful in scale, and genui underserved. But it is not a simple data product play and anyone who approache that way will discover quickly that the incumbents are better positioned to sell p analytics than a startup, and that the government procurement cycle for pure Sa measured in years rather than quarters. The entrepreneur opportunity is a hybri model that uses open source data analytics as the top-of-funnel signal generator wraps it with investigative services, legal partnerships, and outcome-based rever structures that align incentives in ways the incumbents structurally cannot.

The starting point is the data infrastructure itself. Building a pipeline that inge Medicaid spending dataset, joins it against NPPES, the OIG exclusion list, PEC Open Payments, SAM.gov, and state corporate registry data, ar runs it through a set of anomaly detection models is probably a two-to-four pers engineering effort over three to six months to get to a production-grade system. models at this stage do not need to be exotic. The highest-signal fraud indicator this dataset are detectable with relatively straightforward approaches: new entit formation date joined to billing ramp rate, authorized official network graphs

identifying shared principals across multiple NPIs, geographic implausibility sc comparing claimed service delivery locations against population and housing da procedure code billing concentration analysis that flags providers in the top one percent of their taxonomy and state, and temporal pattern analysis identifying th ramp-and-exit signature of bust-out schemes. None of these require a large lang model or a deep learning architecture. They require clean data joins and though feature engineering. The competitive moat at this stage is not the algorithm but data assembly and the domain expertise to know which features actually predict versus which features predict legitimate high-volume providers.

The output of that system is a prioritized list of provider NPIs that warrant investigation, ranked by anomaly score and annotated with the specific signals th triggered the flag. That output is not a fraud finding. It is a fraud candidate list. distinction matters legally, commercially, and ethically. Publishing a ranked list suspicious providers without human investigation and expert review is how start generate defamation lawsuits and regulatory backlash. The data layer generates Humans close them.

The investigative services layer is where the business model gets interesting. The several revenue structures worth considering and a sophisticated entrepreneur s probably pursue multiple simultaneously while the market develops. The first is direct sales to state Medicaid agencies and managed care organizations. State Medicaid directors and MCO medical directors are perpetually under pressure t demonstrate fraud recovery outcomes, and a vendor that can deliver a prioritized investigation queue derived from systematic cross-dataset analysis is solving a re operational problem. The sale is slow, typically twelve to twenty-four months fro first contact to signed contract, and pricing is either a flat annual SaaS fee or a percentage of documented recoveries. The percentage-of-recovery model is more palatable to government procurement because it converts a budget line item into contingency, but it requires the vendor to have the working capital to operate th a long investigation-to-payment cycle before revenue materializes.

The qui tam legal referral model is a second revenue stream that can operate in parallel with the government sales channel and does not require winning procur

contracts. The False Claims Act allows private citizens with original informatio about fraud against the federal government to file suit on the government's beha qui tam relator, receiving fifteen to thirty percent of recovered funds if the government intervenes and the case succeeds. The HCFAC enforcement progran returns $2.80 for every dollar spent on prosecution. A company that systematica identifies high-confidence fraud cases from open source data, builds an evidenti package around each case through investigative fieldwork, and refers those pack to healthcare fraud plaintiff firms on a fee-sharing arrangement is essentially a c origination engine for qui tam litigation. The public disclosure bar in the FCA, limits suits based primarily on publicly available information, requires careful navigation, but a company that combines public data analysis with original field including caregiver interviews, site visits, and beneficiary contact creates origina information that can survive that bar. The legal partnership structure works bec plaintiff firms have the litigation infrastructure but are bottlenecked on case origination. A systematic data-driven lead generator that pre-screens cases for strength is genuinely valuable to them.

The investigative fieldwork component is not optional in this model. It is the pa transforms a billing anomaly into an actionable legal referral and it is the part th hardest to automate. Confirming that a home health agency is billing for phanto visits requires actually attempting to reach the beneficiaries supposedly being se visiting the claimed service delivery addresses, interviewing former employees, a reviewing whatever documentation the agency has filed with state licensing authorities. That is labor-intensive work that requires people with investigative healthcare program knowledge, and the ability to conduct interviews in commur settings. The staffing model for this layer looks more like a private investigative than a software company, which is why most pure-tech founders underestimate i most incumbents in the payment integrity space who have the investigative capa lack the technical infrastructure to generate leads systematically.

The third revenue layer is market intelligence and compliance services sold to legitimate industry participants. Home health agencies that are operating legally significant interest in understanding where their billing patterns fall relative to

providers, both because anomalous-looking billing can trigger audits even for legitimate operators and because understanding the competitive billing landscap helps with rate negotiation and contract management. An analytics platform tha gives compliant home health agencies visibility into how their billing looks from fraud-screening perspective, and helps them maintain documentation practices t will survive scrutiny, is a defensible SaaS product with recurring revenue and a b who is motivated by risk management rather than government procurement. The data infrastructure that generates fraud leads for enforcement purposes generat peer benchmarking and compliance tools for legitimate operators.

The flywheel that makes this business compound over time is labeled outcome d Every investigation that results in a confirmed fraud finding, a successful qui tar recovery, or a state Medicaid audit action creates a labeled data point that impro the model's ability to distinguish true fraud from billing anomalies with innocen explanations. A company that has been running this system for three years has a labeled dataset of confirmed fraud patterns that no one else can replicate from t public data alone, because the public data tells you who billed anomalously but n which of those anomalies turned out to be fraud. That labeled dataset is the actu moat, and it accumulates automatically as the investigative operation generates outcomes. The company that runs this flywheel fastest in a specific taxonomy, wh that is home health, behavioral health, or DME, builds a compounding advantag that space that is genuinely hard for either incumbents or new entrants to replic quickly.

The realistic near-term revenue model for a company in this space, if executed w a combination of three to five state Medicaid agency analytics contracts in the $5 to $2M annual range, a pipeline of qui tam referrals generating contingency fees a two to four year litigation cycle, and a compliance SaaS product for legitimate health operators generating $5K to $20K annually per customer with relatively l churn. None of those revenue streams is enormous in year one. All of them are growing, all of them are somewhat defensible, and the combination of a governm contract channel with a litigation referral channel with a commercial SaaS chann

creates the kind of revenue diversification that makes the business survivable th the slow procurement cycles that characterize this market.

The political tailwind from the current administration's focus on Medicaid frau real but should not be over-indexed. DOGE's involvement in the data release cre short-term attention that benefits companies in this space by elevating the probl the minds of state Medicaid directors, MCO executives, and Congressional overs staff. That attention is useful for accelerating sales conversations that would otherwise take longer to initiate. But the underlying fraud problem predates thi administration by decades and will persist through multiple political cycles. The companies that will win in this space are the ones building durable detection infrastructure rather than chasing the current news cycle, because the news cycl moves on and the structural fraud vulnerability in Medicaid home health does n

← Previous

Nex

Discussion about this post