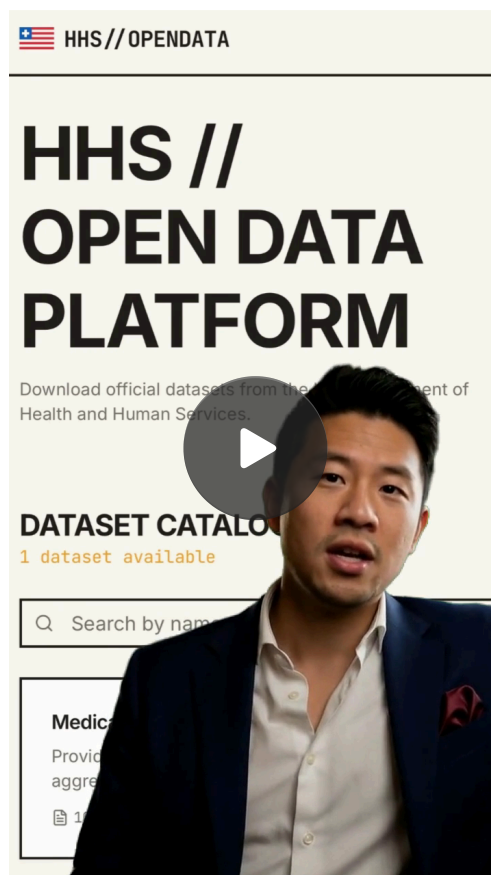


How DOGE Open Sourcing the T-MSIS Medicaid Spend File Unleashed a Distributed Internet Fraud Hunt That Outran Fed Program Integrity, What t Sleuths Found & What Commercial Payers Could Build

MAY 25, 2026 • PAID



Video Preview



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

How DOGE Open Sourcing the T-MSIS Medicaid Spend File Unleashed a Distributed Internet Fraud Hunt That Outran Fed Program Integrity, What the Sleuths Found & What Commercial Paye Could Build

THOUGHTS ON HEALTHCARE • 1:38 PM



Thoughts on Healthcare Markets & Technology is a reader-supported publication. To receive n posts and support my work, consider becoming a free or paid subscriber.

[Listen now →](#)



Thoughts on Healthcare Markets & Technology

 **How DOGE Open Sourcing the T-MSIS Medicaid Spend File Unleashed a Distributed Internet Fraud Hunt That Outran Fed Program Integrity, What the Sleuths Found & What Commercial Payers Could Build**

Thoughts on Healthcare Markets & Technology is a reader-supported publication. To receive new posts and support my work, consider becoming a free or paid subscriber...

▶ Listen now

9 minutes ago · Thoughts on Healthcare

To listen to paid episodes in Apple or Spotify, link your Substack subscription and show settings on those platforms (instructions inside the Substack app under Subscriptions → Podcast).

Thoughts on Healthcare Markets &
Technology is a reader-supported
publication. To receive new posts and
support my work, consider becoming a
free or paid subscriber.

Abstract

1. Feb 13, 2026: DOGE HHS dropped a 10.32 GB Medicaid claims file at opendata.hhs.gov. 227M rows. 1.8M NPIs. Every outpatient and professional HCPCS code from Jan 2018 to Dec 2024. FFS, managed care, and CHIP. Cell suppressed at <12 claims.
2. Within hours, the internet was running notebooks, choropleths, and Substack posts. Within days, identified billing patterns mapped to active DOJ cases. A week later, the file had been mirrored on Hugging Face, Kaggle, and torrents.
3. The federal page is now “under construction.” The data is everywhere. Federal program integrity has been doing this work for 15+ years; the difference is that the public is now in the loop and operating at internet speed.
4. The sleuths surfaced: autism diagnosis mill patterns in MN (mapping to live EIDBI cases), DMEPOS storefronts in TX/FL, telehealth phantom visits, hospital ineligibility patterns, behavioral health overbilling in MD/NJ, and an NY

managed care anomaly cluster that one analyst sized at ~\$90B (unverified, al certainly inflated, but directionally consistent with state-level audits).

5. Federal validation flowed back fast: MN had Feeding Our Future (\$250M, 70% charged), Housing Stabilization Services (\$302M alleged), and EIDBI raids on Smart Therapy and Star Autism. By May 2026, DOJ added 15 trial attorneys explicitly for Medicaid fraud and ran a 15-defendant MN takedown.
6. The commercial side question: \$1.5T in commercial spend, \$30-60B/year of recoverable improper payments, no current cross-payer signal infrastructure, a private payer coalition replicate the open-source pattern compliantly?
7. Short answer: yes. HIPAA permits provider-NPI-aggregated non-PHI public DOJ/FTC Statement 6 gives antitrust safe harbor with 5+ contributors, 3-month lag, 25% cell cap. False Claims Act doesn't apply commercially but contractual bounty structures work.
8. Business model: payer subscription + recovery share + sleuth bounty pool. T clean room + reputation system + tip workflow + LLM pre-triage. Year 3 AR realistically \$30-60M. Defensibility: coalition lock-in plus sleuth network effect.
9. Kill modes: antitrust drift, defamation tail, provider lobby counter-attack, false positive flood, payer retreat after first lawsuit. None fatal if engineered for.

Table of Content

1. Friday the 13th and the file that broke the dam
2. What the data actually is and what it isn't
3. What the sleuths actually did in the first 90 days
4. The federal catch-up and what it tells you
5. Why distributed discovery beats centralized investigation at the top of the f
6. The commercial payer version
7. Legal scaffolding and the HIPAA puzzle
8. Tech stack of a private sleuth community

9. Business model, unit economics, and how money moves
10. Reality check, edge cases, and how this dies
11. Whether this works

Friday the 13th and the file that broke the dam

The drop happened on a Friday the 13th in February 2026, which is the kind of detail the DOGE comms team probably enjoyed. One tweet from the HHS DOGE account, one link to opendata.hhs.gov, and the largest Medicaid claims dataset C has ever published went live with no DUA, no privacy board review, no application form, no waiting period. Treasury Secretary Bessent went on TV the same day reminding everyone that whistleblowers can collect 10-30% of fines under existing federal statute. Bill Ackman quoted that on X and said “Let’s go.” Elon, predictably called DOGE “a state of mind” rather than a department. By Sunday there were heatmaps. By the following Friday opendata.hhs.gov was throttled and the page eventually went under maintenance. As of this writing it still reads “temporarily unavailable while we make improvements.”

The horse is gone. The file has been mirrored on Hugging Face, indexed on Kaggle, at least three independent uploaders, torrented in standard distribution forums, replicated across dozens of academic and amateur data archives. There is no putting this data back in the box.

What sat behind the link was a 10.32 GB CSV-ish artifact: 227M rows, 1.8M unique NPIs across billing and servicing provider fields, every outpatient and professional HCPCS code, monthly grain from Jan 2018 to Dec 2024, total claim count, total payment amount, unique beneficiary count, cell-suppressed at the 12-claim floor. Fee-for-service, managed care, and CHIP all in one file. Sourced from T-MSIS, the Transformed Medicaid Statistical Information System that CMS has been collecting monthly from all 50 states, DC, and the territories for the better part of a decade.

Before Friday the 13th, the closest a non-government analyst could get to this granularity was the T-MSIS Analytic Files via a ResDAC application, which typically chewed through six to twelve months of an academic researcher's life and was effectively unavailable to commercial operators. The pre-existing public T-MSIS releases were de-identified, state-level aggregates, useful for trend lines and not else. Provider-level NPI granularity for the entire program for free, with no friction was new.

What the data actually is and what it isn't



Continue reading this post for free, courtesy of Thoughts on Healthcare.

[Claim my free post](#)

[Or purchase a paid subscription.](#)

[← Previous](#)