Setup Leaves Public Traces: A Public-Data Linkage Playbook for Medicaid Program Integrity

Abstract

State Medicaid program-integrity teams face large provider universes, fragmented oversight responsibilities, and limited investigative capacity. Public records cannot establish Medicaid fraud by themselves, but they can improve the triage process that precedes subpoena, data-use agreement, chart review, and beneficiary-level investigation. This paper develops a public-data linkage playbook for Medicaid fraud, waste, and abuse screening. The framework distinguishes between public setup traces, such as enrollment records, addresses, officers, sanctions, litigation, ownership, public-company disclosures, and site records, and restricted records needed to validate claim execution, such as claims, electronic visit verification, authorizations, payroll, bank records, charts, and beneficiary timelines. Using a New York Medicaid demonstration across six provider categories, the paper shows how public data can convert broad outlier queues into related-entity maps, signal domains, confidence-tiered linkages, and targeted restricted-data requests. In a 928-NPI home-care subset, the workflow narrowed a 441-provider public baseline queue representing $49.4B in payment exposure to a 66-provider public-linkage priority queue representing $8.5B. These figures are not fraud estimates or validated overpayments. They illustrate an operating model for pre-investigation triage, not a causal estimate, fraud finding, or claims-classification benchmark.

Key Takeaways

- Public records cannot prove Medicaid fraud, but they can make pre-investigation triage more reproducible and more precise.
- A useful public-data workflow separates setup traces from restricted execution records and assigns explicit linkage-confidence tiers.
- The New York demonstration supports a playbook for record requests and lead prioritization, not a fraud allegation or loss estimate.

1. Introduction

Medicaid program integrity is a permanent triage problem. State Medicaid Fraud Control Unit directors, state Medicaid agency program-integrity leads, plan special investigations units, and federal partners at the HHS Office of Inspector General and CMS Center for Program Integrity must identify fraud, waste, and abuse across large provider universes, but most leads cannot immediately receive full claims review, chart review, payroll analysis, beneficiary interviews, or financial investigation. The practical problem is not only whether a suspect pattern exists. It is how to decide which leads deserve scarce inves-

tigative attention and which restricted records should be requested first. The question is sharper now: the FY 2025 HHS-OIG Medicaid Fraud Control Unit annual report documents 5,991 managed-care referrals against an enforcement portfolio of nearly $2 billion in combined criminal and civil recoveries [@hhsoigMFCU2025], and ongoing scrutiny of post-PHE Medicaid eligibility redeterminations has reopened federal attention to program-integrity capacity at state agencies [@khodakarami2025UnwindingMedicaidEnrollment].

Federal oversight guidance already frames fraud risk management as an iterative governance function rather than a single enforcement event. The Government Accountability Office describes fraud-risk management as a cycle of assessment, prevention, detection, response, monitoring, and adaptation [@gaoFraudFramework2015]. CMS similarly describes Medicaid program integrity as a state-federal effort that uses audits, education, guidance, data, and technical assistance to combat provider fraud, waste, and abuse [@cmsProgramIntegrity2026; @cmsMedicaidIntegrityProgram2024]. CMS's Center for Program Integrity also emphasizes collaboration with states and the use of predictive analytics [@cmsCPI2026].

The operational pressure is especially clear in managed care. Federal Medicaid managed-care rules require contracts with MCOs, PIHPs, and PAHPs to include procedures for detecting and preventing fraud, waste, and abuse, routine monitoring and auditing, service verification, and prompt referral of potential fraud, waste, or abuse to state program-integrity units or MFCUs [@ecfr4386082026]. Yet oversight reports continue to find uneven referral activity. HHS-OIG reported that some Medicaid managed-care plans made no provider fraud, waste, or abuse referrals in 2022, and that more than half of referring plans made two or fewer provider referrals per 10,000 enrollees [@hhsoigManagedCareReferrals2025]. HHS-OIG's FY 2025 MFCU annual report also shows the scale of the enforcement setting: 53 MFCUs conducted investigations and prosecutions, reported nearly $2 billion in combined criminal and civil recoveries, and received 5,991 fraud referrals from managed-care entities [@hhsoigMFCU2025].

This paper argues that public data should be treated as a formal pre-investigation layer in that ecosystem. The argument is deliberately modest. Public records cannot prove phantom visits, kickbacks, wage underpayment, same-beneficiary duplicate billing, or validated damages. Those questions require restricted records. But public records can identify setup traces: provider identity, enrollment posture, locations, corporate entities, officers, sanctions, public-company relationships, litigation, property records, and other external evidence. Used carefully, those traces can turn broad outlier queues into more specific lead packets and record requests.

The paper contributes a playbook rather than a causal estimate. It combines program-integrity governance, claims-fraud analytics, record-linkage discipline, and public-source data governance into an operational framework for OIGs, SIUs, MFCUs, state Medicaid agencies, and consultants supporting those entities. The New York demonstration is included to show feasibility and to make

the workflow concrete. The paper is not a case study of New York fraud. It is a generalizable method paper with a New York demonstration.

2. Background And Gap

The health-care fraud analytics literature has grown substantially, but much of it focuses on claims data, machine learning models, and audit targeting. Joudaki and colleagues review data-mining studies for health-care fraud and abuse and emphasize that analytic methods can help payers narrow large volumes of claims or claimants for further assessment [@joudaki2015DataMiningHealthcareFraud]. A more recent systematic review by du Preez and colleagues finds a claims-centered machine-learning literature shaped by supervised and unsupervised models, deep learning, scarce labels, privacy constraints, inconsistent data, and a lack of benchmark standardization [@duPreez2025ClaimsFraudMLReview]. Shekhar, Leder-Luis, and Akoglu provide a strong adjacent example: they use unsupervised and explainable machine learning to target hospitals for possible overbilling audits and validate model output against DOJ anti-fraud lawsuits [@shekhar2026MLTargetFraud].

That work is important, but it often starts after restricted claims data are available. The problem addressed here is upstream. Before a state agency, MFCU, plan SIU, or external reviewer obtains detailed claims, EVV, payroll, chart, or bank records, public sources may already contain enough setup information to improve triage. The relevant question is not "can public data classify fraud?" It is "can public data make the next restricted-data request more targeted and defensible?"

Adjacent work in *Health Affairs Scholar* has shown the value of disciplined administrative-data analyses for Medicaid policy questions. Schpero, Zhang, and Civelek use T-MSIS analytic files to characterize Medicaid expansion enrollees in service of the One Big Beautiful Bill Act work-reporting debate [@schpero2025DiagnosedConditionsMedicaidExpansion], and Khodakarami and colleagues use facility-level interrupted time series to show post-PHE unwinding's effect on uninsured emergency department visits [@khodakarami2025UnwindingMedicaidEnrollment]. Both papers illustrate the field's expectation that program-integrity-adjacent quantitative work be tightly framed against a specific decision-maker question.

The record-linkage literature supplies the needed discipline. Fellegi and Sunter's classic framework treats record linkage as a probabilistic problem with explicit matched, unmatched, and uncertain zones [@fellegiSunter1969RecordLinkage]. AHRQ's health-services data-linkage guide emphasizes feasibility assessment, data cleaning, linkage method selection, validation, privacy, and security [@dusetzina2014LinkingData]. More recent health-data linkage work shows that missing data and field selection materially affect probabilistic linkage performance [@li2022DataAdaptiveFellegiSunter]. These lessons are directly relevant to public program-integrity work: a shared address, common phone,

similar legal name, or officer match should not be treated as equal to an exact NPI sanction match.

Responsible analytics guidance also matters. NIST's AI Risk Management Framework gives a useful vocabulary for governing, mapping, measuring, and managing analytic risk [@nistAIrmf2023]. The White House Blueprint for an AI Bill of Rights is nonbinding, but its emphasis on safe and effective systems, data privacy, notice, explanation, and human alternatives is relevant to analytic processes that can affect providers [@ostpAIBillOfRights2022]. In this setting, responsible analytics means not only protecting privacy, but also refusing to turn weak public linkages into unsupported allegations.

The gap, therefore, is practical and methodological. Program-integrity guidance says agencies should detect, refer, and investigate. Fraud analytics literature says data can prioritize audits. Record-linkage literature says messy identifiers require disciplined match rules. Public datasets such as NPPES, LEIE, Open Payments, state corporate registries, and public enforcement records are available [@cmsNPPES2026; @hhsoigLEIE2026; @cmsOpenPayments2026]. What is missing is a practitioner-facing playbook that combines these elements into a reproducible pre-investigation workflow.

3. Conceptual Model: Setup Leaves Public Traces

The playbook begins from a simple distinction: setup leaves public traces; execution leaves restricted traces.

Setup includes incorporating entities, obtaining NPIs, enrolling with Medicaid or Medicare, filing addresses, listing officers, registering business names, occupying sites, contracting with plans, acquiring facilities, appearing in sanctions lists, disclosing ownership interests, and sometimes appearing in litigation or public-company filings. Much of this activity leaves public or semi-public records. Public data is strongest at this stage because it can show whether an organization exists, where it says it operates, which identities it shares with other organizations, whether it appears in exclusion lists, and whether external public records create corroborating context.

Execution includes billing, service delivery, aide presence, authorizations, care plans, payroll, financial flows, patient encounters, EVV records, chart contents, and beneficiary timelines. These records are generally not public. They are held by state Medicaid agencies, MCOs, providers, FIs, payroll processors, banks, or law-enforcement entities. Public data can suggest which execution questions should be asked, but it cannot answer them.

This distinction clarifies the evidentiary role of the playbook. Public linkage is not a substitute for investigation. It is a way to make investigation more targeted. A public-data lead should end with a precise restricted-data request: which NPI, which entity, which time period, which beneficiaries or claims, which payroll or EVV records, and which validation hypothesis. Figure 1 summarizes

this workflow.

Figure 1. Public traces vs private validation workflow

Figure 1: Public traces vs private validation workflow

Note: This figure compares estimates across groups or specifications for the public traces vs private validation workflow. It is intended to make effect heterogeneity and subgroup precision easier to assess.

4. Playbook Methods

The proposed workflow has eight steps.

First, define the provider universe. The unit of screening should be explicit: state, service category, time window, provider identifier, and payment threshold. A state program-integrity unit might begin with all home-care, adult day, NEMT, DME, or SNF providers above a payment threshold. The key is to avoid mixing exploratory source collection with changing denominators. Every provider in the study universe should be traceable to a reproducible inclusion rule.

Second, normalize identity fields. The NPI is the primary key where available, but public linkage also requires normalized legal names, doing-business-as names, addresses, phone numbers, taxonomies, officers, corporate identifiers, EINs, CCNs, and source-specific identifiers. Name and address normalization should be documented before matching begins. Linkage failures and ambiguous matches should remain visible rather than silently dropped.

Third, build a public source inventory. Table 1 lists the source domains used in the demonstration: provider identity, payment exposure, sanctions, corporate records, property and site records, public parent or subsidiary records, nonprofit governance records, litigation and enforcement sources, and SNF ownership and staffing disclosures. Each source should be documented with join keys, refresh cadence where known, evidentiary use, and limitations.

Table 1. Public source domains, join keys, uses, and limitations

Public data domain	Example sources	Program-integrity use	Key limitation
Provider identity	NPPES; NY Medicaid enrollment files	Define provider universe, normalize identities, and detect official-record discrepancies.	NPI issuance does not validate licensure, credentialing, ownership, or service delivery.
Payment exposure	HHS Medicaid Provider Spending	Build cohort, measure public spending surface, and identify peer outliers.	Public spending is a universe or queue measure, not a loss estimate.
Sanctions and exclusions	HHS-OIG LEIE; NY OMIG exclusions	Add high-confidence external contradiction when exact NPI or otherwise high-confidence match exists.	Name-only or fuzzy matches require identity confirmation before narrative use.
Corporate records	NY DOS; other state corporation registries	Map shell entities, related entities, officers, and common-control hypotheses.	Many ownership/member fields require FOIL or paid sources; public extracts may be incomplete.
Property and site records	ACRIS; NYC DOB; OATH; 311; PLUTO	Assess operating-site plausibility and real-estate/control context.	Site complaints or property links are contextual signals, not proof of billing fraud.
Public parent and subsidiary records	SEC EDGAR	Identify public-company relationships, disclosed litigation, and parent-level risk context.	Raw text search produces false positives; use curated and denoised hits only.
Nonprofit governance	IRS 990; ProPublica Nonprofit Explorer	Map governance, officers, related-party transactions, and revenue context.	IRS 990 timing lags and role labels may not match current operational control.
Litigation and enforcement	CourtListener; DOJ/OAG/OSC press releases; settlement/AOD texts	Validate known enforcement examples and identify external-evidence signals.	Raw docket hits must be adjudicated; settlement duties require document-specific review.
SNF ownership and staffing	CMS SNF All Owners; PBJ; HCRIS	Analyze chain ownership, staffing scenarios, and disclosure patterns.	SNF per-diem and staffing-gap outputs are scenarios unless validated against payroll, rosters, and claims.

Notes: This table documents the source files, scripts, variables, or data inputs used in the analysis. It is included to make the construction of the analytic evidence reproducible.

Fourth, link records using explicit confidence tiers. Table 2 gives the core linkage structure. Exact NPI matches across NPPES, state enrollment, LEIE, or OMIG

are high-confidence. EIN exact matches are high-confidence when available. Legal-name exact matches after normalization are medium-high because names can change or collide. Exact address or phone matches are medium-confidence and should never stand alone. Officer or signatory matches are medium-high but require human review when names are common or inconsistent.

Table 2. Linkage confidence tiers and signal domains

Rule or domain	Confidence or cap	Data inputs	Permitted use
NPI exact match	HIGH	NPPES; state enrollment; LEIE; OMIG	Sanction overlap, cohort intake, revalidation status.
EIN exact match	HIGH	IRS 990; NPPES other identifiers	Owner or tax-entity linkage.
Legal-name exact match after normalization	MEDIUM-HIGH	NPPES; NY DOS; LEIE; OMIG	Sanction near-match or corporate crosswalk when NPI is absent.
USPS-normalized address exact match	MEDIUM	NPPES; NY DOS; ACRIS	Address-cluster signal when combined with another signal.
Phone exact match	MEDIUM	NPPES; NY DOS; manual web capture	Phone-hub signal when combined with another signal.
Officer or signatory exact name match	MEDIUM-HIGH	NY DOS; IRS 990; ACRIS	Common-control inference after human review.
Network and identity	within-group cap 2	address cluster; phone cluster; dual-channel pairing; enhanced flag	Identify related-entity clusters and structure review packets.
Billing anomalies	within-group cap 2	fast ramp-up; abrupt stop; PPP workforce gap; baseline volume and top-decile paid counted jointly	Prioritize broad billing outliers for restricted-data review.
Regulatory contradictions	within-group cap 3	sanction; not in enrollment; revalidation overdue; multistate exclusion	Identify compliance contradictions and state-side validation needs.
External evidence	manual/external evidence domain	CourtListener; SEC EDGAR; OMIG exact-NPI; DOJ/OAG/OSC sources where curated	Corroborate public signals with enforcement, litigation, or public-company disclosures.
Medium-confidence compounding rule	medium signals require corroboration	all linkage domains	Prevent false-positive escalation from weak public links.

Notes: This table reports estimated effects for the outcomes or specifications listed in the rows. Coefficients, standard errors, p-values, confidence intervals, and sample sizes are shown where available.

Fifth, group signals by independent evidence domains. The demonstration uses four broad domains: network and identity, billing anomalies, regulatory contradictions, and external evidence. This grouping prevents correlated signals from

being overcounted. For example, an address cluster and phone cluster may both be part of the same underlying organizational structure; counting them as fully independent can inflate confidence. By contrast, a high-confidence sanction match and a separate public-company disclosure may represent more independent evidence.

Sixth, apply anti-false-positive rules. Medium-confidence signals cannot stand alone. Raw text-search hits are not evidence until adjudicated. Fuzzy name matches are research leads, not facts. Public spending exposure is not loss. Public linkage outputs should be labeled as signals, candidates, queues, or validation prompts.

Seventh, create lead packets. A lead packet should not read like an accusation. It should state the provider universe, public signals, linkage confidence, alternative explanations, source limitations, and the exact restricted records needed next. A useful packet helps an investigator decide what to request, not what to conclude.

Eighth, maintain provenance. Every count, dollar figure, source extract, and linkage rule should be traceable to a source artifact, generator script, snapshot date, or manual extraction note. In the demonstration, each source artifact is preserved with a source locator and a documented manuscript use.

5. New York Demonstration

The demonstration uses a New York Medicaid public-data project as a feasibility test for the playbook. The source project constructed a six-category cohort of 3,563 distinct NPIs across CDPAP/home-care, social adult day care, LHCSA plus CHHA, non-emergency medical transportation, durable medical equipment, and SNF-family providers. The cohort is based on 2018-2024 HHS Medicaid Provider Spending and more than 30 public enrichment sources. The cohort-visible Medicaid payment surface is approximately $91B in HCPCS-visible public spending. This number is a pattern-universe measure, not a fraud estimate or loss estimate.

Table 3 summarizes the demonstration cohort. The important feature is not the absolute dollar amount. It is the ability to define a reproducible provider universe and then layer public source domains onto that universe without changing denominators midstream.

Table 3. New York demonstration cohort

Scope or category	NPI count	Spending or amount	Evidence bucket	Caveat
Six-category New York Medicaid cohort	3563	approximately $91B HCPCS-visible Medicaid paid, 2018-2024	B1 pattern universe	Not a fraud estimate, loss estimate, harm figure, or validated overpayment.
Unique non-SNF NPIs across CDPAP/home-care, SADC, LHCSA+CHHA, NEMT, and DME	2441		B1 pattern universe	Five non-SNF category counts sum to 3,072 before 631 cross-category overlaps are removed.
SNF-family raw category count	1145	$16.5B HCPCS-visible layer reported separately	B1 pattern universe / scenario context	SNF true per-diem layer is substantially higher; staffing-gap scenario is illustrative, not measured overpayment.
SNF-family unique after non-SNF overlap removal	1122		B1 pattern universe	Raw SNF-family count includes 23 NPIs also present in non-SNF buckets.
CDPAP / home-care subset used for baseline precision-filter demonstration	928	baseline-flagged $49.4B; not-baseline-flagged $32.9B	B2 queue and unflagged remainder	Comparison has not been rerun on full six-category cohort.
Social adult day care category	505	$4.66B paid	B1 pattern universe	Category payment surfaces may overlap and should not be summed as additive loss.
LHCSA plus CHHA category	417		B1 pattern universe	Aide-level service-line validation requires restricted records.
Non-emergency medical transportation category	779	$2.91B paid	B1 pattern universe	Trip-level details vary by MCO; public data cannot validate trip delivery.

Scope or category	NPI count	Spending or amount	Evidence bucket	Caveat
Durable medical equipment category	443	$1.31B paid	B1 pattern universe	DME has a distinct Medicare-enrollment expectation; public Medicare-crosswalk signal is especially interpretable here.

Notes: This table summarizes policy timing, cohorts, thresholds, or state-level sample construction. It is intended to make the identifying variation and comparison groups transparent.

The cleanest quantitative example is the 928-NPI CDPAP/home-care subset. In that subset, a four-rule public baseline screen flagged 441 providers, representing $49.4B in payment exposure. The public-linkage layer then identified 66 providers within that baseline queue, representing $8.5B in payment exposure, using the limited enhanced signals implemented for that comparison. This is an 85.0% provider reduction and 82.8% dollar reduction within the baseline-flagged queue.

This result should be read narrowly. The comparison has not been rerun on the full six-category cohort. The measured baseline is not a full SIU rule stack with authorizations, charts, payroll, EVV, or beneficiary interviews. The public-linkage layer in the comparison is a precision filter within the baseline queue, not an independent recall test. The enhanced-only cell is zero by construction because the enhanced review pool was seeded from the baseline queue. It is therefore not evidence that public linkage finds no providers missed by baseline screening.

Table 4. Home-care public-linkage precision-filter demonstration

Metric	Providers	Paid amount	Evidence bucket	Interpretation
Cohort used for comparison	928		B1 pattern universe	Defines the measured comparison universe.
Baseline public screen flagged providers	441	$49.4B	B2 administratively unsupported queue	Baseline public rules create a broad queue that needs triage.
Baseline plus public-linkage flagged providers	66	$8.5B	B3 public-data proxy queue	Public linkage functions as a precision filter within the baseline queue.
Baseline flagged but not public-linkage flagged	375	$40.9B	B2 remainder	These providers remain broad public-screen leads rather than public-linkage packet leads.
Public-linkage flagged but not baseline flagged	0 structural		structural artifact	Do not interpret as recall evidence.
Neither baseline nor public-linkage flagged	487	$32.9B	unprioritized public remainder	Remainder shows the limit of public-tier methods.
Reduction from baseline queue to both-flagged queue	85.0% reduction	82.8% reduction	precision-filter statistic	Headline result of the home-care precision-filter demonstration; within the baseline-flagged queue only.

Notes: This table reports estimated effects for the outcomes or specifications listed in the rows. Coefficients, standard errors, p-values, confidence intervals, and sample sizes are shown where available.

Even with these caveats, the demonstration is useful. It shows how public linkage changes the operational question. A broad public outlier queue asks, "Which providers are statistically unusual?" A public-linkage packet asks, "Which unusual providers also have public setup traces that justify specific restricted-data review?" The second question is more useful for an OIG, SIU, MFCU, or MCO deciding how to spend investigative capacity.

6. Validation And Guardrails

Validation in a public-data playbook should be signal-specific. It should not imply that every method surfaces every known fraud case. In the New York demonstration, 13 public enforcement cases were identified within the 3,563-NPI cohort. All 13 were in the cohort. The Medicare-crosswalk "Medicaid-

only" filter fired on 12 of 13, with a falsification-style correct negative for a non-emergency-medical-transportation case that legitimately appeared in Medicare and therefore should not have triggered the Medicaid-only signal. LEIE/OMIG exact-NPI sanction overlays fired on 5 of 13. Curated EDGAR fired on 3 of 13. Maildrop or PMB heuristics fired on 3 of 13. The cluster-2+ signal fired on only 2 of 13. PPP and dual-channel proxy signals fired on 0 of 13.

Table 5. Known enforcement-case validation summary

Validation signal	Cases	Denominator	Direction	Caveat
Known enforcement case appears in 3,563-NPI cohort	13	13	positive coverage	Does not mean every method surfaced every case.
Medicare-crosswalk Medicaid-only filter	12	13	strongest signal coverage	The one negative, V-13 Mobile Life Support, is a correct negative because ambulance/NEMT can legitimately bill Medicare.
LEIE/OMIG exact-NPI sanction overlay	5	13	high-confidence external corroboration	Only exact-NPI matches should be treated as high-confidence without further identity confirmation.
Curated EDGAR signal	3	13	selective external-evidence coverage	Use curated/denoised EDGAR hits only; raw text hits can overproduce false positives.
Maildrop / apartment / PMB heuristic	3	13	weak-to-moderate identity/site signal	Maildrop alone cannot support a narrative claim.
Cluster with at least two public-linkage signals	2	13	limited known-case recall	Cluster-2+ dominates the home-care precision filter but does not explain most known enforcement cases.
PPP workforce-gap signal	0	13	no coverage in validation set	PPP jobs reported is self-declared and should remain a workforce proxy, not a fraud allegation.

Validation signal	Cases	Denominator	Direction	Caveat
LHCSA/SADC dual-channel proxy	0	13	no coverage in validation set	Requires claim-level same-beneficiary/date validation.
Overall validation interpretation	signal-specific, not uniform recall	13	guardrail	Do not claim the stack uniformly surfaces all thirteen cases.

Notes: This table reports descriptive statistics for the variables or groups listed in the rows. Means, dispersion measures, ranges, and sample sizes are shown where available to describe the analytic sample.

This pattern is informative. It suggests that some public signals generalize better than others. Medicare-crosswalk and sanction overlays were more powerful against known enforcement cases than cluster density, PPP workforce proxies, or dual-channel pairing. That does not make the weaker signals useless. It means they should be framed as hypothesis generators or support signals rather than validation evidence.

The demonstration also uses explicit damages buckets. B1 is the pattern universe: total paid amounts in the cohort. B2 is an administratively unsupported or broad-review queue. B3 is a clinically or operationally unsupported public-data proxy queue. B4 is validated overpayment, which requires settlement, AOD, judicial finding, chart review, EVV, payroll, claims, bank, or beneficiary-level validation. Public data can populate B1, B2, and some B3 proxy queues. It cannot populate B4 except by pointing to already-public enforcement resolutions.

Guardrails are part of the method. A public-data workflow should include negative controls, alternative explanations, source preservation, exact confidence labels, and a rule that medium-confidence linkage cannot stand alone. It should also preserve a "not evidence" category for raw search hits, fuzzy matches, and public facts that are interesting but not probative.

7. Restricted-Data Escalation

The practical endpoint of the playbook is a restricted-data request. Table 6 maps public signal classes to the minimum records required for validation. A phantom-visit hypothesis requires EVV logs, aide schedules, and beneficiary attestation. An unenrolled-NPI hypothesis requires CMS-855 or state enrollment files and credentialing records. A common-control hypothesis requires disclosable-party histories, corporate filings, bank-signatory cards, and lease or landlord records. A kickback hypothesis requires bank tracing, referral logs, and plan-of-care signatory patterns. A same-beneficiary dual-channel hypothesis requires claims or encounter records with beneficiary and service-date detail.

Table 6. Public signal to restricted-data escalation map

Public signal class	Minimum restricted records	Validation purpose
Phantom visits or aide-allocation gap	EVV log; aide schedule; beneficiary attestation interview over at least a 30-day window	Test whether billed services were delivered by available aides to actual beneficiaries.
Unenrolled-NPI billing or enrollment contradiction	CMS-855A/B/I; state enrollment file; provider credentialing file	Confirm enrollment, ownership, disclosed parties, and eligibility to bill.
Upcoding or bundled-service split	Chart sample of at least 30 encounters; corresponding 837P or 837I claims	Test whether coding and bundled-service treatment match records.
Kickback or referral-source concentration	Bank tracing; referral logs; plan-of-care signatory pattern	Test money flow and referral/plan-of-care concentration.
Common-control or shell-entity relationship	CMS-855 disclosable-party history; corporate filings; bank-signatory cards; lease and landlord records	Confirm beneficial owners, managers, and operational control.
SNF staffing gap	PBJ raw quarterly submission; payroll; HR roster; agency-staff invoices	Promote staffing scenario to validated operational finding if supported.
Prior-settlement or AOD compliance breach	AOD compliance file; monitor reports; post-settlement billing	Test continuing obligations, notice filings, and monitored compliance.
Encounter-data anomaly	Full T-MSIS extract for state-year-category; MCO encounter feed	Validate same-beneficiary, same-day, duplicate, or cross-channel hypotheses.

Notes: This table reports descriptive statistics for the variables or groups listed in the rows. Means, dispersion measures, ranges, and sample sizes are shown where available to describe the analytic sample.

This structure helps prevent two common errors. The first is overclaiming: treating a public signal as though it proves an execution-stage fact. The second is underusing public data: dismissing public records because they cannot by themselves prove fraud. The correct middle position is that public data should improve the precision of restricted-data requests.

Figure 2 summarizes the handoff from queue to packet to record request.

Figure 2. From broad queue to restricted-data request

Figure 3 shows the logic of a public linkage graph. Exact NPI sanctions can be high-confidence corroboration. Shared addresses, shared phones, and officer matches are useful, but they remain medium-confidence unless corroborated by other evidence domains.

Figure 3. Public linkage graph schematic

8. Operational Implementation

A state or plan implementing this playbook does not need to start with a large machine-learning system. A minimal implementation requires a defined

Figure 2. From broad queue to restricted-data request

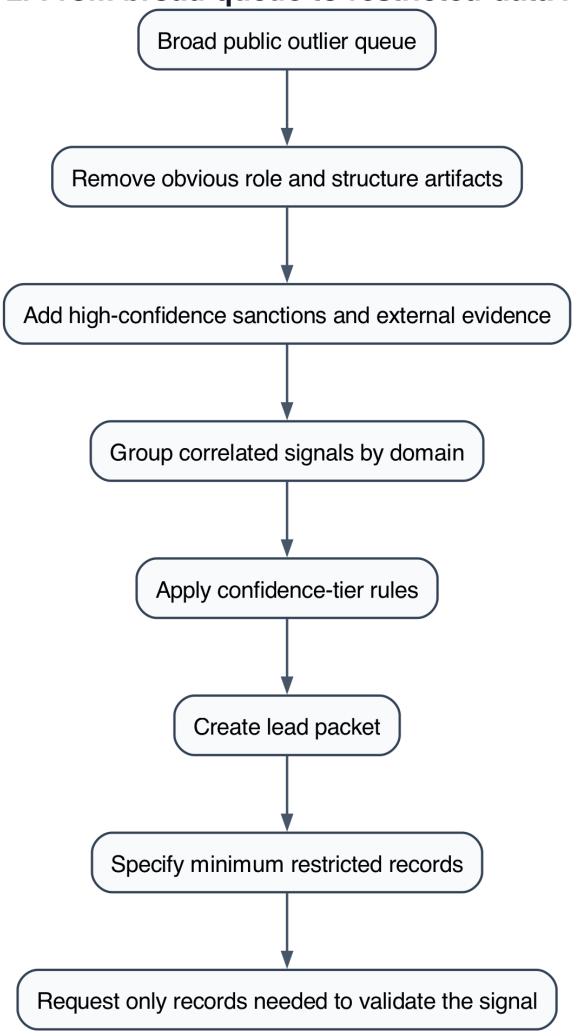

Figure 2: From broad queue to restricted-data request

Note: This figure provides contextual structure for the from broad queue to restricted-data request. It summarizes the policy setting, mechanism, or empirical workflow used to interpret the estimates.

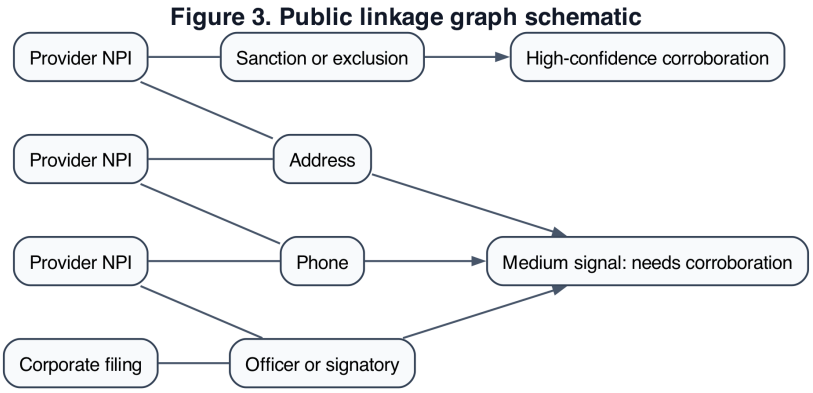

Figure 3: Public linkage graph schematic

Note: This figure compares estimates across groups or specifications for the public linkage graph schematic. It is intended to make effect heterogeneity and subgroup precision easier to assess.

provider universe, a source inventory, reproducible normalization scripts, a linkage-confidence protocol, a review queue, and a packet template.

The first operational decision is scope. A program-integrity team should choose one service category and one time window rather than beginning with all Medicaid. High-risk categories may include home care, adult day services, NEMT, DME, or long-term care, but the right starting point depends on state priorities, data access, and investigative capacity.

The second decision is source governance. Public sources change. NPPES refreshes. Sanctions lists update. State corporate registries and provider enrollment files may alter formatting. A usable workflow should preserve source snapshots or source dates, document refresh cadence, and keep hashes or manifests for derived files.

The third decision is review design. Analysts should not hand investigators a spreadsheet of thousands of flags. They should produce tiered packets that separate high-confidence matches, medium-confidence corroborated patterns, low-confidence research leads, alternative explanations, and the minimum restricted records needed next. Packet quality matters because it determines whether the workflow helps investigators or merely shifts analytic burden downstream.

The fourth decision is human review. Public-data linkage is an assistive process. It should not automatically trigger adverse provider action. A human reviewer should examine exact matches, ambiguous identities, source limitations, and reasonable non-fraud explanations before any referral leaves the analytic team.

The fifth decision is feedback. When restricted-data review confirms, rejects,

or modifies a lead, that outcome should return to the source-linkage workflow. Over time, the team can learn which public signals are strong in a particular service category and which signals generate false positives.

9. Limitations

This paper has several limitations.

First, the New York demonstration is not a causal design and does not estimate the effect of public linkage on fraud recoveries, conviction rates, audit yield, or overpayment detection. It is a methods and feasibility demonstration.

Second, the cleanest quantitative result is limited to a 928-NPI home-care subset. The same baseline-vs-public-linkage comparison has not been rerun across the full 3,563-NPI six-category cohort. The result should be described as a precision filter within a baseline queue, not a general recall estimate.

Third, public data is uneven across service categories. DME may have interpretable Medicare-crosswalk signals because Medicare enrollment is common in that market. NEMT, home care, adult day services, and SNF-related analyses require different source expectations. A signal that is strong in one category may be irrelevant or misleading in another.

Fourth, public-source fields may be self-reported, stale, incomplete, or ambiguous. Addresses can be mailing addresses, practice addresses, administrative addresses, or shared office suites. Legal names can change. Phones can be shared by billing services. Officer names can collide. These limitations make confidence tiers necessary.

Fifth, the demonstration uses known public enforcement cases for directional validation, but that validation set is not a random sample of fraud. Public enforcement cases reflect detection, prosecution, settlement, media, and reporting processes. They are useful for checking whether signals fire on known examples, but they cannot establish population sensitivity or specificity.

Sixth, advanced model outputs from later exploratory work are intentionally excluded from the main evidence base. They may be useful in a future appendix, but only after reproducibility review confirms what each model predicts and whether labels are independent of score construction.

Finally, legal and ethical constraints are central. Public-data outputs can affect reputations and business relationships. Manuscripts, referrals, and internal dashboards should avoid naming non-enforcement targets unless legal review approves a specific use. The method is safest and most generalizable when it reports patterns, source domains, confidence rules, and restricted-data asks rather than allegations.

10. Conclusion

Public records are not a substitute for Medicaid program-integrity investigation. They are a way to make investigation more disciplined. A public-data linkage playbook can help oversight teams define provider universes, normalize identities, join public sources, group independent signal domains, assign confidence tiers, preserve provenance, and generate precise restricted-data requests.

The central claim is practical: public data cannot validate fraud, but it can improve the queue that precedes validation. That is enough to matter. In a system where investigative capacity is limited and program-integrity responsibilities are distributed across states, plans, OIGs, SIUs, MFCUs, and federal partners, better triage is not a side issue. It is the operating layer that determines which questions get asked, which records get requested, and which risks receive human attention.

References

Citation metadata is stored in `literature/bibliography.bib`. This paper uses Pandoc-style citation keys pending journal-specific formatting.

Appendix: Public-Data Medicaid Fraud Playbook

Appendix A. Data Governance

The manuscript uses a preserved set of audited source artifacts. Each public source, derived table, and manual extraction note is documented with a source locator, extraction date when available, and manuscript use. Materials outside the audited evidence base are not used for substantive claims.

Appendix B. Manuscript Tables

Table	Purpose
Table 1	Public source domains, join keys, uses, and limitations.
Table 2	Linkage-confidence rules and grouped signal domains.
Table 3	Demonstration cohort and category surfaces.
Table 4	Home-care precision-filter result.
Table 5	Known enforcement-case validation summary.
Table 6	Public signal to restricted-data request map.

Notes: This table documents the source files, scripts, variables, or data inputs used in the analysis. It is included to make the construction of the analytic evidence reproducible.

All tables are derived from audited artifacts, with extraction rules and rerun triggers documented in the project materials.

Appendix C. Manuscript Figures

Figure	Purpose
Figure 1	Shows the movement from provider universe to public setup traces to restricted-data validation.
Figure 2	Shows how a broad queue becomes a lead packet and record request.
Figure 3	Shows how NPI, address, phone, officer, corporate, and sanction nodes interact.

Notes: This table documents the source files, scripts, variables, or data inputs used in the analysis. It is included to make the construction of the analytic evidence reproducible.

Appendix D. Linkage Confidence Rules

The paper uses the following evidence hierarchy:

Rule	Linkage type	Confidence	Manuscript use
L-01	NPI exact match	High	Sanctions, cohort intake, revalidation, and enrollment checks.
L-02	EIN exact match	High	Tax-entity and owner linkage when EIN is available.
L-03	Legal-name exact match after normalization	Medium-high	Sanction near-match or corporate crosswalk when NPI is absent.
L-04	USPS-normalized address exact match	Medium	Address-cluster signal requiring corroboration.
L-05	Phone exact match	Medium	Phone-hub signal requiring corroboration.
L-06	Officer or signatory exact-name match	Medium-high	Common-control inference after human review.

Notes: This table documents the source files, scripts, variables, or data inputs used in the analysis. It is included to make the construction of the analytic evidence reproducible.

Medium-confidence linkages cannot stand alone in a Tier-1 narrative or manuscript claim. They must be combined with at least one other medium-or-better signal, preferably from a different signal domain.

Appendix E. Signal Domains

The manuscript groups public signals into four domains:

1. Network and identity: address clusters, phone clusters, name roots, officer links, and dual-entity structures.
2. Billing anomalies: public payment outliers, ramp-up patterns, paid-per-claim patterns, abrupt stops, and workforce proxies.
3. Regulatory contradictions: sanctions, exclusions, enrollment discrepancies, revalidation status, and cross-state exclusion checks.
4. External evidence: public enforcement, curated litigation, SEC disclosures, public audit reports, and settlement/AOD records.

The grouping rule is designed to prevent correlated public traces from being counted as independent evidence.

Appendix F. Dollar Buckets

The manuscript uses four dollar buckets:

Bucket	Meaning	Permitted interpretation
B1	Pattern universe	Total public payment surface under study.
B2	Administratively unsupported or broad-review queue	Providers or dollars worth further review.
B3	Public-data proxy queue	Higher-priority public signals requiring restricted-data validation.
B4	Validated overpayment	Claims tested against settlement, AOD, judicial finding, chart review, EVV, payroll, bank, beneficiary, or other restricted evidence.

Notes: This table documents the source files, scripts, variables, or data inputs used in the analysis. It is included to make the construction of the analytic evidence reproducible.

The approximately $91B full-cohort payment surface is B1. The $49.4B baseline-flagged home-care amount is B2. The $8.5B both-flagged home-care amount is B3. Public data does not itself populate B4.

Appendix G. Known-Case Validation

The validation matrix is directional and signal-specific:

Signal	Coverage in 13 known cases	Interpretation
Cohort presence	13 of 13	Known cases are inside the analytic universe.
Medicare-crosswalk Medicaid-only filter	12 of 13	Strongest validation signal; one correct negative for NEMT/ambulance.
LEIE/OMIG exact-NPI sanction overlay	5 of 13	High-confidence external corroboration where exact-NPI match exists.
Curated EDGAR	3 of 13	Useful for public-company relationships and external disclosures.
Maildrop / PMB heuristic	3 of 13	Supporting site/identity signal, not standalone evidence.
Cluster-2+	2 of 13	Useful precision filter in home-care subset, limited known-case recall.
PPP workforce proxy	0 of 13	Exploratory support signal only.
LHCSA/SADC dual-channel proxy	0 of 13	Requires beneficiary-level claims validation.

Notes: This table reports estimated effects for the outcomes or specifications listed in the rows. Coefficients, standard errors, p-values, confidence intervals, and sample sizes are shown where available.

Appendix H. Restricted-Data Escalation

The playbook treats public signals as prompts for restricted-data requests. Examples:

Public signal	Minimum validation records
Phantom visit or aide-allocation gap	EVV logs, aide schedules, beneficiary attestation, and claims.
Enrollment contradiction	CMS-855, state enrollment files, credentialing files, and disclosable-party records.
Upcoding or bundled-service split	Chart sample and corresponding 837P or 837I claims.
Kickback or referral-source concentration	Bank tracing, referral logs, and plan-of-care signatory patterns.
Common-control or shell-entity signal	CMS-855 disclosures, corporate filings, bank-signatory cards, leases, and landlord records.
SNF staffing scenario	PBJ raw submissions, payroll, HR roster, agency-staff invoices, and claims.
Settlement compliance question	AOD files, monitor reports, compliance certifications, and post-settlement billing.
Encounter anomaly	T-MSIS extract or MCO encounter feed with beneficiary-level detail.

Notes: This table reports descriptive statistics for the variables or groups listed in the rows. Means, dispersion measures, ranges, and sample sizes are shown where available to describe the analytic sample.