

RosterProof Methodology

How the data is collected, matched and kept honest. For Medicaid program integrity units, MFCUs and enterprise compliance teams. Figures queried from the published views on 2026-10-20 (feed refreshed 2026-10-20; NPI registry snapshot September 2026).

45 sources: 4 federal, 40 states + DC	341,901 active exclusions (346,591 current records)	56,954 active records with an NPI verified in NPPES	0 entity-type conflicts across 28,671 verified NPIs	Weekly refresh · full history kept (394,471 versions)
--	--	--	--	--

What it is, and is not

One continuously refreshed dataset of **who is excluded or sanctioned, by which authority, since when, and (where we can prove it) which NPI they hold**: the OIG LEIE, SAM.gov, the FDA Debarment List, the FDA Clinical Investigator Disqualification list, 40 state Medicaid exclusion lists and DC. It is the **exclusion and owner-screening layer that feeds compliance, payment-integrity and fraud programs**. It is **not a fraud detector**: it reports published exclusion status and does not judge conduct or score risk.

Sources and collection

Every source is the **official agency's own publication**, never a mirror; the raw file is archived byte-for-byte before parsing so any figure traces to what the agency published that day. By format: spreadsheet 24, CSV/TSV 8, web table 4, JSON endpoint 3, **PDF 6** (loaded only through a reconciliation gate). Delivery exceptions: **NJ** emails its file monthly (filed automatically from the inbox); **MA and NH** block automated retrieval, so a person saves the file monthly into a drop folder; **MI** was released to us under Michigan's public-records law. The pipeline runs **weekly**; NPI matching runs monthly with each NPPES release.

Normalization

Forty-plus layouts become one record: source, jurisdiction, person or organization, name parts, state, reason (verbatim), excluding authority, start and end date, and whether the end date was **published by the agency or observed by us**. A strict **source-fidelity rule** applies: what the agency publishes is what is loaded. Nothing is corrected, reinterpreted or suppressed, even if it looks impossible; unparseable values are left empty **with the literal text preserved**; every duplicate collapse must reconcile numerically and keep the absorbed rows' differences.

NPI matching

Tier	What it means	Active records	Confidence
T1	The agency published the NPI and the registry confirms it and a matching name	31,751	HIGH
T2	No NPI published; exactly one provider in the whole national registry fits the normalized name and state (registered alternate names included)	23,634	HIGH
T3	The state license number agrees with the registry	426	MEDIUM
T4	Several candidates fit; an AI model weighed specialty and license class and wrote its reasons	1,143	LOW (flagged)

Human-confirmation rule: an automated match is never the last word. **needs_human_review** = do not act automatically; have a person verify. It is set on 6,119 active records: 3,826 whose published NPI is now **deactivated**, 1,150 whose published NPI the registry **could not confirm**, and the 1,143 AI-judged matches. A name-only result is never presented as confirmed: T2 is accepted only on a unique fit and states its basis on every record, and the **283,797 active records (83.0%) with no NPI** ship without a guessed one, labelled "name match only".

Quality controls

- **Reconciliation gates:** source rows to loaded records with every difference named; a census proving no published value became empty without a recorded reason; tests run on the exact odd values found.
- **PDF gate (AK, DE, HI, NE, NV, WY):** consistent columns; an **independent second extractor** that must agree row-for-row; the PDF's own printed count; per-column validators; a verbatim spot-check; a sweep proving no page text was dropped. A failed gate stops the load.
- **Change detection and exclusion_end_basis:** each record carries a content hash over its published fields; the weekly comparison classifies every record as ADDED, CHANGED, REMOVED or REINSTATED and only real changes enter the change feed (meaningless fields such as portal IDs and file names are excluded). When a record disappears from an agency's list, its end is dated by us and labelled **EXCLUSION_END_BASIS = REMOVED_BY_SOURCE**; an end date the agency itself published is labelled as published, so a subscriber always knows whether an end is the agency's statement or our observation.
- **Guardrails and audit:** a load outside a source's normal row band, with drifting null rates or new columns is held back and the prior data carries forward; the automated audit currently shows **0 data failures** (the one open item is a listing-metadata reminder, not data).
- **Persistent-failure alarm:** a source failing two consecutive runs is registered as a blocker; its last good data is never silently dropped.

House rules

- **One published truth:** the Snowflake tables, the download files and the API come from the same weekly run and always match.
- **Existing views never change shape;** additions are additive and follow a fixed redeploy checklist, so integrations never break.
- **Non-exclusion statuses (for example a Medicare opt-out) are kept in separate views of the same shape** and never mixed into the exclusion tables, the API, the download files or the change feed, so a screening query cannot return one by accident.
- Every derived value carries a basis column; raw values are never destroyed.

Limits, stated honestly

- **It reports exclusion status; it does not detect fraud.** An exclusion is an agency's published action, not RosterProof's finding.
- **Verify before any adverse action** with the source agency (and OIG's SSN-based search for the federal list); every API response and download carries this notice.
- It can only catch what agencies publish; names and practice states in the registry can lag; a weekly refresh means a new exclusion can take up to a week to appear.
- **No list, no data:** three states (FL, ID, WV) are not yet loaded; twelve jurisdictions (seven states and five territories) publish no list.
- A provider operating under a new organizational NPI, entity and officer is invisible to this method; ownership linkage needs enrollment data only a program holds.

Attribution and use labels

Each source's agency, terms and required notices are at rosterproof.com/attribution and in the README with every snapshot. Federal sources are U.S. Government works in the public domain. **North Carolina (166 active records) and Texas (11,953) are included under those states' non-commercial permissions**, so the data is labelled **for compliance-screening use** and NC/TX rows must not be resold or repackaged commercially without the agencies' written permission. Agencies disclaim the accuracy of their lists; RosterProof passes records through as published.