Contents

Methodology

As of 2026-05-30. Methodology evolves with the corpus and the architecture; revision history is maintained in the project repository.

What Provenant Does

Provenant is an oncology-focused price-transparency data product. It ingests hospital-disclosed negotiated rates for oncology drugs and services, normalizes them across hospitals, and surfaces rate dispersion with per-row, cryptographically signed provenance. It is built for litigation economists, pharma market-access analysts, and health-system network strategists who must defend every figure they cite. What distinguishes Provenant is architectural rather than editorial: every rate row carries signed provenance back to its source file, and every cross-hospital aggregate is recomputable and independently re-verifiable from its contributing packs. We surface dispersion. We do not adjudicate cause.

Data Sources

Provenant draws on four classes of public source data. Each ingested artifact is version-pinned and hash-recorded, so any downstream figure traces to the exact bytes it came from.

Hospital machine-readable files (MRFs). The negotiated-rate disclosures required under the CMS Hospital Price Transparency rule (45 CFR 180). We ingest a hospital's MRF when it is published or refreshed. Each row inherits the file's SHA-256 (`mrf_sha256`), `source_url`, and `fetch_timestamp`. We do not edit, impute, or smooth disclosed rates. Our fetcher honors `robots.txt` without exception; where a hospital's federally-mandated file sits behind a blanket CDN block, we record the gap rather than bypass it.

Medicare ASP files. CMS publishes Part B Average Sales Price files quarterly. Each ASP value is pinned to the quarter in force for the rate's effective period — 2026Q2 across the current corpus — and recorded per row. An expert recomputing a figure must pull that quarter's file, not whatever ASP file is current at read time.

HCRIS cost report data. We use the CMS-derived HCRIS summary for one signal: receipt of IPPS Disproportionate-Share (DSH) payments, the recomputable input to 340B DSH-route eligibility. Pinned to release vintage.

CMS HCPCS file and CMS ASP NDC-HCPCS crosswalk. We use the CMS HCPCS file to resolve a billed J/Q code to its drug descriptor, and the CMS-published quarterly ASP NDC-HCPCS Crosswalk to attach the corresponding HCPCS-level Medicare Part B ASP. We do not resolve a J-code to an individual NDC for class purposes: the published Medicare ASP is itself CMS's volume-weighted average across the NDCs under a HCPCS code, and we adopt that HCPCS-level

basis rather than reconstruct a per-NDC price. Where a HCPCS code is retired or replaced, each row remains pinned to the code and ASP-quarter in force at `fetch_timestamp`, so a figure is reproducible against the vintage it was computed on rather than silently re-mapped. Drug-class refinement against the FDA NDC database — replacing the current brand-default classification with originator-vs-biosimilar-vs-generic resolution — is a roadmap item (ADR 0036); until it lands, no per-NDC class is asserted, and every drug row carries `markup_assumption: brand-default` explicitly.

We deliberately do not use three sources. We do not use **claims data**: it is licensed and patient-level, and tempts causal inference Provenant does not make; disclosed rates suffice to surface dispersion. We do not query **HRSA OPAIS** as a live source: it exposes no machine-readable endpoint suitable for reproducible ingestion, so we resolve 340B eligibility structurally rather than read enrollment. We do not yet ingest **payer Transparency-in-Coverage files**: that work is bot-gated and structurally blocked on a payer-access layer, and sits on the roadmap.

The Two-Tier Pack Architecture

Every Provenant dataset ships as a signed pack carrying one of two tiers. The tier is a claim about evidentiary strength, not completeness. The model is documented in ADR 0031.

Litigation-grade. The source MRF is CMS JSON (v3.0); it passes the official CMS Hospital Price Transparency validator (`cms_validator_pass: true`); and a CY2026 attestation is captured from a named hospital officer, with date and CMS Certification Number (CCN). Inclusion also requires source freshness and minimum oncology coverage. Where a pack misses a criterion, the miss is disclosed rather than hidden — Cedars-Sinai is retained at a 165-day source age with the recency deviation recorded on the pack itself.

Analytics. The source is a non-JSON MRF — CSV, or JSON inside an archive — for which the CMS validator is not yet wired. These packs carry the same provenance discipline (SHA-256, write-once storage, signed Ed25519 manifest, full `verify-pack` pass) but record `cms_validator_pass: null` with a stated reason ("validator not wired for csv format") and `attester_name: null`. The field is three-state: passed, failed, or not-evaluated. A `null` must never be read as a `false`; an un-run validator is not a failed one.

Roll-up packs derive their tier and never silently promote. A roll-up is litigation-grade only if every contributing source row is litigation-grade; otherwise it is analytics. Tier flows up from the weakest contributing source, by construction. Promoting analytics sources to litigation-grade — by wiring the CMS validator to CSV and to archived JSON — is a roadmap item that would re-tier existing packs without re-fetching.

Per-Row Provenance and Caveat Discipline

Provenance is per row, not per file. Every rate row carries, at minimum: `mrf_sha256` (the hash-pinned source file), `source_url`, `fetch_timestamp`, `code_type` (HCPCS, CPT, or REV), `payer_canonical_id`, `asp_unit_caveat`, `route_340b`, and `contributing_tier`. A reader can take any single row and walk it back to the exact disclosed bytes that produced it.

Caveats are composable, not hierarchical. Each row carries a `caveats` array, so distinct caveats stack rather than overwrite. We do not collapse them into a single confidence score, because that would hide which specific limitation applies to a given figure. A row may be both a modeled-340B-markup row and a unit-basis-suspect row, and both labels travel with it.

Three caveats deserve description.

The **`asp_unit_basis`** caveat fires when ASP is below $1 per unit, or when the MRF's `drug_unit` field is missing. In practice it fires on essentially every drug J-code, because MRFs rarely disclose a unit basis. We state this plainly: for most drug rows, any comparison between a disclosed rate and an ASP reference is made without a hospital-disclosed unit basis, and the caveat marks exactly

that. A high raw ratio computed across unreconciled units is reported as a flagged candidate, never as a clean headline.

The **brand-default assumption** is applied when a drug's FDA NDC class is unknown. Rather than guess generic-versus-brand, we default to brand and label the row (`markup_assumption: brand-default`), so the assumption is auditable rather than silent. Biosimilar Q-codes currently take the brand default, flagged, pending FDA NDC-class refinement (roadmap).

The **`modeled_340b_markup`** label marks any computed markup proxy as modeled, never as an observed or quoted price. Each markup row also carries `markup_basis` ("modeled"), `route_340b` (the eligibility basis), and a composed `markup_basis_full` string, so a reader sees the full assumption stack on the row without cross-referencing this page.

The 340B Methodology

This is the section most likely to be attacked, so it is written most conservatively. The decisions are documented in ADRs 0035 and 0036.

Eligibility, not enrollment. Provenant resolves 340B *eligibility* — a structural, statute-grounded property — and never claims *enrollment*, an administrative status in HRSA's OPAIS registry. OPAIS exposes no machine-readable endpoint suitable for reproducible verification from a clean checkout, so active enrollment is surfaced only as provisional, hand-confirmed context (`enrollment_confirmed`) and is never an analytical gate. This is a documented access constraint, not a bypassed one.

Two parallel routes. Eligibility is resolved by either of two independent statutory routes:

- **DSH route.** The hospital receives IPPS Disproportionate-Share payments (HCRIS `Total_DSH_Payments` > 0). At the payment magnitudes observed across this corpus ($8–19M), receipt of the DSH payment adjustment is a sound, recomputable proxy for a DSH adjustment percentage above the §340B threshold of 11.75% (PHSA §340B(a)(4)(L)). The exact adjustment percentage (HCRIS Worksheet E Part A line 4.03) is a documented refinement we report as such; the binary eligibility determination here rests on receipt of the payment, labeled as a proxy rather than as the percentage itself.
- **Cancer-hospital route.** The hospital is a PPS-exempt cancer hospital under 42 CFR 412.23(d). These hospitals are not paid under IPPS, so their HCRIS DSH is $0; they qualify via this route instead. Membership is read from a configuration file enumerating the statutorily-fixed set; it is hand-curated but citable — each entry is tied to the regulation and CCN-verified against CMS enrollment data, the opposite of a fuzzy guess.

The eligibility flag is the disjunction: `is_340b_eligible = DSH-route OR cancer-route`. It is never derived from enrollment. The two routes are parallel eligibility classes, not a hierarchy; a hospital may qualify under both (`route_340b: "dsh+cancer"`), and the route is labeled on every applicable row.

The two-route design is conservative-by-design for a measured reason. Four of the seven corpus hospitals receive $0 IPPS DSH because they are PPS-exempt cancer hospitals. A DSH-only test would have wrongly zeroed the markup on exactly the hospitals carrying the headline drug findings.

The markup proxy. The true 340B ceiling is AMP − URA, and AMP is confidential. So where a markup is reported, it is modeled off public ASP and labeled `modeled_340b_markup`:

```
markup_340b_proxy = negotiated_rate − (medicare_asp × (1 − d))
```

where `d` is the statutory minimum discount by drug class — 0.231 for brand/innovator (the 0.769 brand-default factor), 0.171 for clotting-factor and exclusively-pediatric drugs, 0.13 for generic. It is a floor proxy, not a ceiling price, and is never represented as one.

Because `d` is the statutory minimum discount, `medicare_asp × (1 − d)` is the highest floor the statute permits — so the modeled markup is the smallest markup consistent with 340B pricing,

not a maximal one. The true ceiling (AMP − URA) reflects unit-rebate adjustments, including inflation penalties, that generally deepen the discount below the statutory minimum; the real markup is therefore at least as large as the modeled figure. We additionally use public ASP as a proxy for the confidential AMP, and these can diverge; the `modeled_340b_markup` label marks both simplifications. We report the conservative direction by design.

Under 42 CFR Part 10, 340B ceiling prices are confidential; Provenant does not publish, quote, or reconstruct them.

To be explicit: an eligible hospital is one that *could* participate by statute. Provenant makes no claim that any hospital is enrolled, is purchasing at 340B prices, or is acting improperly. Eligibility is a structural fact; conduct is not something Provenant asserts.

Cross-Hospital Aggregation

Rate dispersion is meaningful only once payers are comparable across hospitals. Provenant resolves payer identity with a deterministic, hand-curated canonical lookup (ADR 0032): roughly twelve canonical entries collapse the national spine and the next-most-frequent insurers, and the long tail stays passthrough. Matching is boundary-aware prefix matching — never substring, never fuzzy. The discipline is deliberate: "First Health Network" contains the substring "health net," but boundary-aware matching keeps it distinct from "Health Net," and the thirty-plus independent Blue Cross Blue Shield licensees are not merged, since that would assert a false cross-hospital equivalence. Cross-hospital comparison is made only for canonical payers; a passthrough rate is never silently compared against a canonical one.

This canonicalization is intentionally narrow: it resolves approximately 37% of normalized rows; the remaining ~63% — regional and niche plans — stay passthrough and are never collapsed. We treat this as a coverage boundary, not a defect: a smaller set of defensible cross-hospital comparisons is worth more in an expert report than a larger set built on guessed payer equivalences.

For each `(canonical_payer, code)` tuple, Provenant computes `hospital_count`, `rate_min`, `rate_p10`, `rate_p50`, `rate_p90`, `rate_max`, and `rate_max_over_min` (the dispersion ratio). Every aggregated row cites its inputs: a `source_pack_ids` field lists the contributing source packs, and a per-hospital vector maps each contributing CCN to its rate, so any cell expands back to the specific signed source rows that produced it. The source packs are themselves signed under named filter views — `headline` (rows where conflict_flag AND outlier_flag are both set on an active code: the strongest single-row claims), `conflict_pairs` (cross-payer rate conflicts on the same tuple), and `all_oncology` (unfiltered width: every normalized oncology row the source MRF emits). The roll-up reconstructs only from rows its signed sources contain. Views are defined as Polars expressions in `src/oncorate/affidavit/views.py`; predicates are not operator-supplied, by deliberate restriction documented in ADR 0037. The aggregation invariant is documented in ADR 0033.

Within-hospital aggregation. A hospital's MRF may publish more than one rate for what reduces to a single (canonical_payer, code) tuple — from plan-tier differences, MRF generation artifacts, or unsurfaced modifier variation. Every disclosed rate is preserved as its own provenance-bearing row in the source pack: the source pack is the lossless record, and no rate is collapsed or hidden at parse time. For the cross-hospital roll-up, each hospital's contribution to a tuple is reduced to its representative rate via the **median across that hospital's preserved rates for that tuple**. The median is deterministic, reproducible from the source pack, robust to within-MRF outliers, and chosen explicitly over min, max, or most-recent rules (each of which would introduce a directional bias into the markup analytic). The selection rule is documented in `src/oncorate/affidavit/builder.py`; a reader expanding any roll-up row back to its source packs sees both the per-hospital representative rate and the underlying set it was reduced from.

Across-hospital aggregation. The roll-up's `rate_min`, `rate_p10`, `rate_p50`, `rate_p90`, `rate_max`, and `rate_max_over_min` are computed across the participating hospitals' representative rates (one per hospital per tuple). These statistics are descriptive of the contributors observed, not estimates of any population parameter; at low `hospital_count` (3–4), the `rate_p50` is a median over three or

four values and should be read as a midpoint of the observed contributors, not as a robust central tendency.

Row counts vary by source-MRF structure, not by pipeline behavior. Because the pipeline never imputes rows, the number of rows a hospital contributes is a property of how that hospital filed its MRF. The ratio of unfiltered (`all_oncology`) rows to filtered (`headline`) rows is a structural diagnostic, not a pipeline defect. A hospital that files one rate per code with little per-payer differentiation yields almost no within-hospital conflicts and a near-empty conflict view; a hospital that files across 100+ payers makes nearly every oncology row a cross-payer conflict, so its conflict and unfiltered views nearly coincide; a conflict-dense bare-wide C-code MRF widens only slightly. We treat these divergences as disclosed source structure (ADR 0037), surfaced before any analysis runs — not as inconsistency to be normalized away.

Recursive Verification

Every pack is signed with Ed25519, and the public key ships with the pack; verification needs no Provenant-held secret. A roll-up pack carries a `source_packs` field — one entry per contributing pack, each recording that pack's `pack_id`, hospital CCN, tier, and `manifest_sha256` — and the whole set is covered by the roll-up's signature. The `oncorate-verify-pack` tool re-verifies every contributing source pack before it verifies the roll-up: it recomputes each source pack's manifest SHA-256, asserts it matches the recorded value, and recursively verifies that source pack in full. Verification is therefore transitive: a passing roll-up asserts that every pack beneath it also passes, down to the original MRF SHAs in each source pack's provenance records. It is not a request to trust the aggregation step. Alter any source pack after aggregation and its manifest SHA no longer matches; the `source_packs` check flips to FAIL and the roll-up cannot verify.

Reproducibility has two layers, and we distinguish them. First, the signed parsed rates in any pack re-verify offline from the pack alone: the manifest binds each file's SHA-256 and the Ed25519 signature covers the manifest, so the figures are tamper-evident without network access and without any Provenant-held secret. Second, the row-level provenance records the source MRF's own SHA-256 and `fetch_timestamp`, so any figure proves which version of the hospital's file produced it. Upstream re-publication does not alter or invalidate a prior signed pack — the pack is self-contained and byte-pinned. Raw source MRFs are retained in a content-addressable write-once store keyed by SHA-256 (`fetch.storage.GcsWormStore` / `LocalWormStore`); under the production deployment of that store, the original parse is re-derivable from the raw bytes by hash. Production bucket provisioning is operational follow-up (ADR 0034); until then, raw-byte re-derivation depends on retention by the publishing hospital or by Provenant's local archive. We do not assert that any pack is "most-recent-possible," only that it is the hospital's published MRF at `fetch_timestamp`, byte-pinned by SHA-256.

What We Don't Do

We do not ingest claims data. We do not publish, quote, or reconstruct 340B ceiling prices, which are confidential under 42 CFR Part 10. We do not infer a drug's NDC class — we apply a labeled brand-default assumption, pending FDA NDC-class refinement. We do not gate analytics on enrollment status; eligibility is resolved structurally and enrollment is never claimed. We do not claim to cover all 340B-eligible hospitals. And we do not adjudicate causation, fraud, or impropriety. We surface dispersion; experts adjudicate.

Current Coverage

Provenant is seven hospitals deep: four litigation-grade (Stanford, Cedars-Sinai, MSK, Fred Hutchinson) and three analytics-tier (UMass Memorial, Roswell Park, MD Anderson). The current cross-hospital UnitedHealthcare roll-up spans six of the seven — Roswell Park's MRF discloses HCPCS and DRG codes but does not publish UnitedHealthcare contracts at the

canonical-spine level, so it is in the corpus but does not contribute to the UHC roll-up. (Source-MRF disclosure scope determines which rollups a hospital can contribute to; this is a property of the hospital's published file, not a pipeline filter.) The UHC roll-up contains 278 aggregated (canonical_payer, code) rows as of 2026-05-30; of these, 41 reach four-hospital depth on a single tuple and 71 reach three-hospital depth, with the remainder at one- or two-hospital depth. Several additional cancer-center CSVs are seeded but await a resumable large-file fetch path (roadmap). Payer-side Transparency-in-Coverage ingestion is on the roadmap, structurally blocked on a payer-access layer rather than on methodology. Coverage grows on a customer-funded model: hospitals are added as engagements require, not speculatively.

How to Verify Any Claim

Any figure on this site traces to a signed pack, and any pack re-verifies from a clean checkout in one command:

```
oncorate-verify-pack --pubkey signing_public_key.pem <pack-directory>
```

The public key (`signing_public_key.pem`) ships inside every pack. For a cross-hospital roll-up, the tool re-verifies each contributing source pack first, then the roll-up, terminating in a per-check report:

```
verify: reports/2026-05-30-allviews-rollup/rollup_pack_2026-05-30_cross_hospital_rollup_uhc
  TIER: analytics
  manifest            PASS
  README.md           PASS
  cover.pdf           PASS
  rates.csv           PASS
  rates.json          PASS
  pack_content        PASS
  manifest_signature  PASS
  pack_metadata       PASS
  tier_metadata       PASS
  source_packs        PASS
all checks PASS (10/10)
```

The ten checks: **manifest** (present and well-formed); **README.md / cover.pdf / rates.csv / rates.json** (each present and SHA-256-matched to the manifest's file map); **pack_content** (the content hash binding the file set — and, for a roll-up, the source-pack set — matches the value embedded in cover.pdf); **manifest_signature** (the Ed25519 signature over the manifest verifies against the shipped key); **pack_metadata** (the manifest carries `pack_version` and `pipeline_version`, recording the schema generation and the build version of the pipeline that produced this pack); **tier_metadata** (the manifest carries a valid tier label — litigation-grade, analytics, or explicitly legacy-unset — matching the three-state semantics applied throughout: a null is never read as a false, and tier is never silently promoted at build time); and **source_packs** (each contributing pack's manifest SHA is recomputed, matched, and recursively re-verified). A litigation expert can run this from a clean checkout in a single command, with no Provenant-held secret and no network access to Provenant. Verification time scales with the size of the contributing packs and the recursion depth of the source-pack chain.