

# How We Learned What We Didn't Know

---

## Discovering and Measuring CDM Gaps Through AI-Augmented Downstream Validation

---

Technical White Paper

September 2026

Roger Carlson

Senior Data Analytics Engineer

Corewell Health Research Institute, Grand Rapids, MI

Sam Martin

Director, Data Architecture

City of Hope Cancer Center, Los Angeles, California

---

### Summary

---

Our OMOP CDM passed every quality check the community offers. The All of Us Research Program accepted our export. Data Quality Dashboard was green across the board. We had no reason to think anything was missing.

Then I tried to use it. I picked a simple honest broker project — dog bites at our children's hospital ED — and tried to reproduce it from OMOP instead of Caboodle. The original analyst had found 1,245 patients. I got 687. That was the first failure. We tried eight more completed studies. All of them failed. Zero of nine achieved 90% match against the original source results.

The question that drove everything after that point: what would it take to use OMOP the way we use Caboodle — as a viable data source for unique, ad-hoc research projects?

At Corewell Health (22 hospitals, 4.3 million active patients, single Epic instance), we developed three discovery techniques to answer that question and applied them using AI-augmented development tooling. The result: 71 table-level gaps, 90.4 million vocabulary routing rows recovered, a visit-gatekeeper cascade that unlocked 270 million orphaned clinical records, and a CDM that grew from 3.78 billion to 8.42 billion rows over five months. A single analyst executed the entire methodology.

This paper describes what we found, how we found it, and what happened when we closed the gaps.

---

### Introduction

---

We built our OMOP CDM for All of Us. That was the original mandate — satisfy the AoU data quality requirements and deliver a conformant export. We did. The CDM passed DQD. It passed AoU's validation

pipeline. By every measure the OHDSI community provides, the CDM was correct.

But I work in honest broker analytics. My team pulls data for researchers — ad-hoc cohorts, custom datasets, variables that don't map neatly to standard study designs. We do that work from Epic Caboodle. The question I wanted to answer was simple: could we do that work from OMOP instead?

I picked the easiest project I could find. A completed study on dog bites at our children's hospital ED. The original analyst had queried Caboodle and found 1,245 patients. I wrote the equivalent OMOP query and got 687.

The 558 patients who were missing had encounters in departments the CDM didn't cover, coded with visit types the vocabulary mapping didn't include, linked through table relationships the ETL had never been taught to traverse. Every one of them would have passed DQD if they'd been present. They weren't.

I tried eight more projects. They all failed. A hernia repair study at 51% match. A shoulder arthroplasty study at 0% — zero patients found. A ventral hernia study at 3%. The CDM wasn't inaccurate. It was incomplete. The data it contained was correct. The data it was missing was enormous.

The question stopped being "is the CDM good enough for AoU" and became "what would it take for OMOP to work the way Caboodle works — as a data source for any research question a PI walks in with?"

Everything that follows was driven by that question.

---

## Methods

---

### The Enabling Technology

This work was executed using Snowflake Cortex Code, an AI-augmented development environment that generates SQL iteratively, classifies files at scale, and searches source table metadata semantically. A single analyst operated the tooling. The AI did not make decisions. It generated candidate SQL, ran validation queries, and classified files. Every clinical judgment, vocabulary mapping decision, and architectural choice was human.

The distinction matters. The claim is not that AI built our CDM. The claim is that AI made a single person productive enough to do work that previously required a team and years of calendar time.

### Discovery Technique 1: Archaeological Scan

We scanned 3,650 files from institutional honest broker archives on network storage. These files represent real research questions that real analysts answered by querying Epic source systems directly.

A PowerShell pipeline copied the files locally. A Python uploader staged the SQL content to a Snowflake table. Snowflake Cortex (mistral-large2) performed two-pass triage: the first pass flagged files referencing novel source tables, the second pass filtered for files that would teach an experienced analyst something new. The automated triage reduced 3,650 files to 502 requiring human evaluation — an 86% reduction.

From the triage output, we extracted every source table reference from FROM and JOIN clauses and classified each against a CDM coverage map. A “gap” was a source table that appeared in research queries but had no corresponding ETL pipeline in the CDM.

**Result: 71 table-level gaps across 315 verified research projects.**

The gaps were not uniformly important. DateDim (a utility table for date arithmetic) appeared in 214 of 315 projects but represents a structural convenience, not clinical data. ProcedureEventFact appeared in 121 projects and represented genuine missing clinical procedures. We prioritized by cross-project frequency weighted by clinical significance.

| Priority            | Gaps | Projects Affected | Resolution Rate |
|---------------------|------|-------------------|-----------------|
| P0 (near-universal) | 6    | 50-214 each       | 100%            |
| P1 (high demand)    | 20   | 10-50 each        | 100%            |
| P2 (medium demand)  | 14   | 5-10 each         | 93%             |
| P3 (specialized)    | 12   | <5 each           | 67%             |

Each gap routed to a fix layer: STCM entry (vocabulary mapping), ETL pipeline (new dbt model chain), vocabulary routing model (cross-domain correction), or documented structural limitation (no CDM analog exists).

## Discovery Technique 2: Fitness Validation

Nine completed research datasets were selected for full reproduction from OMOP alone. The selection criteria: the original project must have produced a concrete patient count or dataset from Epic source data, the clinical domain must overlap with existing CDM content, and the original analyst’s query must be available for comparison.

For each project, we classified every variable in the delivered dataset: available in OMOP, available with an ETL fix, available with an STCM addition, available with a vocabulary routing correction, or a systemic gap with no current CDM analog.

**Result: 0 of 9 projects achieved 90% match at baseline.**

| Project                   | Source Result | OMOP Result | Match |
|---------------------------|---------------|-------------|-------|
| Dog Bites (HDVCH ED)      | 1,245 pts     | 687 pts     | 55%   |
| Hernia Repair (Blodgett)  | 986 pts       | 502 pts     | 51%   |
| Shoulder Arthroplasty     | 179 pts       | 0 pts       | 0%    |
| Ventral Hernia (Blodgett) | 655 pts       | 19 pts      | 3%    |
| Anxiety and C-Section     | 2,095 pts     | ~7,892 pts  | Over  |
| Cirrhosis LVP             | 298 pts       | 432 pts     | 145%  |
| Pancreatectomy            | 108 pts       | 11 pts      | 10%   |
| Colorectal Cancer         | 3,080 pts     | 1,934 pts   | 63%   |
| ASD-PAH (system-wide)     | 153 pts       | 70 pts      | 46%   |

Two projects were not feasible from OMOP at any level of ETL investment (0-10% match). One over-counted because the CDM lacked obstetric exclusion tables that the original analyst used. Six under-counted by 37-55%.

Root causes split cleanly:

| Category                                                | Type        | Projects |
|---------------------------------------------------------|-------------|----------|
| Location granularity (facility-only, no department)     | Systemic    | 6        |
| Procedure mapping incomplete (CPTs not in ETL source)   | ETL-fixable | 4        |
| Visit/encounter cardinality mismatch                    | Hybrid      | 2        |
| Missing clinical domain tables (OB, flowsheets)         | Systemic    | 2        |
| Vocabulary routing (billing CPTs broader than surgical) | ETL-fixable | 1        |

Each failure generated diagnostic patterns stored in a searchable knowledge base. Subsequent validation attempts reused those patterns and avoided rediscovering the same problems. The knowledge base grew from 0 to 293 patterns across 14 clinical domains during this phase.

### Discovery Technique 3: Cross-Domain Vocabulary Routing

This discovery came from a single failed validation.

A validated honest broker query returned 1,302 patients from Caboodle for a cardiac study. The same query against OMOP returned fewer. Patient-level comparison revealed the gap: ICD-10 codes that had been diagnosed in these patients existed in the source system but not in the CDM's `CONDITION_OCCURRENCE` table. The codes were valid. The vocabulary mapping was correct. The records had been silently dropped during ETL routing.

The mechanism: OMOP convention routes records to the table matching the standard concept's DOMAIN\_ID, not the source concept's domain. An ICD-10 code in the Condition domain that maps to a standard concept in the Observation domain belongs in OBSERVATION, not CONDITION\_OCCURRENCE. Standard ETL patterns route by source domain and never check whether the standard concept landed elsewhere.

We queried CONCEPT\_RELATIONSHIP to find every source-to-standard mapping where the source domain differs from the target domain. The query is simple:

```
SELECT source.DOMAIN_ID AS source_domain,
       target.DOMAIN_ID AS target_domain,
       COUNT(*) AS affected_mappings
FROM CONCEPT source
JOIN CONCEPT_RELATIONSHIP cr ON source.CONCEPT_ID = cr.CONCEPT_ID_1
JOIN CONCEPT target ON cr.CONCEPT_ID_2 = target.CONCEPT_ID
WHERE cr.RELATIONSHIP_ID = 'Maps to'
      AND source.DOMAIN_ID != target.DOMAIN_ID
GROUP BY source.DOMAIN_ID, target.DOMAIN_ID
ORDER BY affected_mappings DESC
```

Cross-referencing against our ETL routing logic quantified the actual data loss:

| Source Domain to Target Domain    | Dropped Rows | Affected Patients |
|-----------------------------------|--------------|-------------------|
| ICD-10 Condition to Observation   | 9,600,003    | 1,304,062         |
| SNOMED Observation to Procedure   | 68,518,813   | 2,772,317         |
| SNOMED Observation to Measurement | 5,472,630    | 973,623           |
| ICD-9 Condition to Observation    | 4,283,892    | 477,326           |
| ICD-10 Condition to Measurement   | 2,548,590    | 103,039           |
| Total                             | 90,424,000   | ~2,800,000        |

90.4 million rows. 2.8 million unique patients. Silently absent.

Six vocabulary routing models were built to recover these rows. Any OMOP site can reproduce the detection query in an afternoon.

---

## Results: What Happened When We Closed the Gaps

---

Discovery without resolution is an academic exercise. We closed the gaps.

### Phase 1: Pipeline Construction (Weeks 1-4)

Six new ETL pipelines were built from gap analysis, each following the established dbt PULL/STAGE/WRITE pattern:

| New Source                   | Target CDM Table     | Rows Produced |
|------------------------------|----------------------|---------------|
| ProcedureEventFact           | PROCEDURE_OCCURRENCE | 45,294,168    |
| ImagingFact                  | PROCEDURE_OCCURRENCE | 9,625,657     |
| ReferralFact                 | OBSERVATION          | 10,446,335    |
| CoverageDim                  | PAYER_PLAN_PERIOD    | 6,041,268     |
| AnesthesiaRegistryMetricFact | MEASUREMENT          | 3,938,348     |
| IcuStayRegistryDatamart      | MEASUREMENT          | 192,139       |

## Phase 2: Visit Cascade (Week 5)

The visit-as-gatekeeper architecture amplifies gaps nonlinearly. Every clinical domain STAGE model in our ETL uses an INNER JOIN to VISIT\_OCCURRENCE. If a patient's encounter is not in the visit table, every clinical event attached to that encounter is silently excluded from every domain.

We identified 15 ambulatory encounter types flagged as face-to-face in Epic but absent from the CDM's visit type vocabulary mapping. Adding them produced 17.9 million new visits.

The cascade: those 17.9 million visits unlocked 270 million previously orphaned clinical records across all domains. A 15:1 amplification ratio.

| Domain               | Gain                | % Increase   |
|----------------------|---------------------|--------------|
| MEASUREMENT          | +204,152,888        | +8.1%        |
| CONDITION_OCCURRENCE | +36,060,128         | +8.1%        |
| NOTE                 | +9,134,819          | +3.3%        |
| OBSERVATION          | +4,681,980          | +4.9%        |
| PROCEDURE_OCCURRENCE | +1,016,522          | +1.6%        |
| DRUG_EXPOSURE        | +1,050,665          | +0.9%        |
| <b>Total cascade</b> | <b>+270,381,495</b> | <b>+7.0%</b> |

Build time decreased from 1h 8m to 59m despite adding 270 million rows — a clustering optimization (PERSON\_ID on MEASUREMENT\_RAW) offset the volume growth.

## Phase 3: SDOH Expansion (Week 5)

Five Social Determinants of Health pipelines deployed, mapping 98 STCM entries across SocialDeterminantFact, SmartData social history flowsheets, and PatientExtendDimX.

OBSERVATION grew from 100.6 million to 350.2 million rows (+248%). SDOH was the single largest untapped clinical domain in the CDM.

## Phase 4: Vocabulary Resolution (Weeks 6-20, ongoing)

The STCM pipeline replaced Usagi entirely. 28,857 source codes discovered across 43 vocabulary domains. 97.9% resolved. 10,430 production STCM rows. 6,127 custom concepts with CONCEPT\_ID >= 2,000,000,000. Zero Usagi involvement. (This work is described in detail in the companion poster, “Beyond Usagi.”)

## Phase 5: Continued Pipeline Expansion (Weeks 6-20)

Surgical device exposure (instruments and implants via SurgicalSupplyDim), episode expansion (TransplantEpisodeFact, cancer staging), and ICD-10-PCS vocabulary integration. The CDM continued growing as each new pipeline unlocked source data that had been invisible.

## Cumulative Impact

| Metric                  | Baseline (April 2026) | Current (September 2026) |
|-------------------------|-----------------------|--------------------------|
| CDM rows                | 3,780,000,000         | 8,420,000,000            |
| Growth factor           | —                     | 2.2x                     |
| Patients                | 3,780,000             | 7,400,000                |
| OMOP domains populated  | 14                    | 14                       |
| dbt models              | 274                   | 320+                     |
| STCM rows               | ~3,500                | 10,430                   |
| Custom 2B+ concepts     | 0                     | 6,127                    |
| Knowledge patterns      | 0                     | 787                      |
| Build time (production) | 1h 8m                 | ~1h 10m                  |

The CDM grew 2.2x in five months. Build time barely changed.

## Generalizable Findings

### 1. Cross-domain vocabulary routing gaps are likely widespread

The detection query runs against standard OMOP vocabulary tables that every site has. If your ETL filters by source domain before joining to standard concepts, you may be silently dropping rows where the standard concept resides in a different domain. The query takes five minutes to write and an afternoon to analyze.

We found 90.4 million dropped rows. We would be surprised if we were the only site with this pattern.

## 2. Visit-as-gatekeeper amplifies gaps nonlinearly

Adding 17.9 million visits produced a 270 million row cascade (15:1). Any site using INNER JOINs to visit\_occurrence in staging logic should audit visit completeness before pursuing domain-specific gap analysis. The visit table is the structural backbone. If it is incomplete, everything downstream is incomplete by the same proportion.

## 3. DQD compliance is not research readiness

A perfect DQD score means the CDM conforms to structural constraints. It does not mean the CDM contains the data researchers actually need. The nine fitness validation projects demonstrated this concretely: DQD-compliant CDM, 0 of 9 projects reproducible at 90%.

Sites should not conflate structural quality with coverage completeness. They measure different things.

## 4. AI-augmented development makes this methodology reproducible

The archaeological scan classified 3,650 files in hours. The fitness validation accumulated 787 diagnostic patterns across 200+ projects. The pipeline construction built six new ETL chains in a single development sprint. A single analyst executed the entire methodology.

The resource requirement drops from a multi-person team to one domain expert with AI tooling. The methodology is not novel in concept — every CDM builder knows gaps exist. What changed is the economics. Discovery and resolution that was prohibitively expensive became continuous work.

---

## Sustaining the Improvement: A Closed-Loop Architecture

---

Discovery and resolution are not enough. A CDM that improves once is a project. A CDM that improves continuously is infrastructure. After closing the initial gaps, we faced the harder question: how do you prevent regression and ensure new knowledge feeds forward?

We built five interlocking components.

### Component 1: Knowledge Base

Every time a validation fails, the failure teaches something. A join path that silently drops rows. A filter that excludes unexpected encounter types. A dimension table where deleted records reappear unless explicitly excluded.

These lessons have a half-life of zero if they live in the analyst's memory. **ACQUIRE\_KNOWLEDGE** is a Snowflake table with one row per validated pattern — domain classification, problem description, and example SQL. The table is indexed via Snowflake Cortex Search for semantic retrieval.

The knowledge base grew from 0 to 787 patterns across 88 domains. Early validation batches hit 6 new patterns per 8 projects. Recent batches hit 1-2. The declining rate is the convergence signal: we are encountering fewer novel problems per project.

Before any clinical query, the analyst (or the AI agent) must search the knowledge base. We call this "ACQUIRE Pre-Flight." The protocol is mandatory, not advisory.

## **Component 2: Blind Reproduction Protocol**

Given a completed research project delivered from Epic source data, an AI agent generates a reproduction query from OMOP alone — without seeing the original SQL. The result is compared at the patient level. Discrepancies are diagnosed. Each diagnosis either confirms the CDM or reveals a gap.

Over 200 projects validated across 10 batches. Pattern discovery per batch is declining, confirming convergence.

## **Component 3: Curation Layer**

109 automated quality checks implemented as a dbt macro + CSV seed pattern. Adding a check means adding a row to the CSV. The curation layer produces a QA\_LOG (violation counts per build) and QA\_EXCLUSIONS (row-level exclusion list for FATAL violations).

When first deployed: 41 million FATAL violations. After systematic resolution: 23,000. A 99.9% reduction. The 41 million were primarily duplicates (19 million), null person\_ids, and temporal validity failures. The curation layer is portable — the macro and seed file copy to any dbt project with OMOP tables.

## **Component 4: Source-to-Concept Mapping Pipeline**

The standard community tool for STCM authoring is Usagi. We replaced it entirely with a four-tier pipeline running within Snowflake: auto-disposition, deterministic SQL, LLM with source data context, and clinical note extraction. 28,857 source codes discovered across 43 vocabulary domains. 97.9% resolved. 10,430 STCM rows. 6,127 custom concepts. (Described in detail in the companion poster, "Beyond Usagi.")

## **Component 5: AI Agent with Ontological Routing**

CARE\_RESEARCH\_AGENT is a Snowflake Cortex Agent with 59 tools. When a clinical question arrives, the agent searches the knowledge base (Pre-Flight), looks up the ontology to determine which semantic views cover the clinical domain, generates SQL, and validates against known patterns.

The agent is both consumer and contributor. When it encounters a query it cannot answer correctly, the failure is diagnosed, a pattern is inserted into the knowledge base, and the next query benefits.

## **The Closed Loop**

Each component feeds the others:

- Research demand discovers gaps (reproduction protocol)
- Gaps route to fix layers (ETL pipeline, STCM entry, curation check, or knowledge pattern)
- Fixes are validated against the original demand
- The knowledge base grows with every cycle

The system converges toward source equivalence with every research question that pushes against its boundaries. The CDM is not a snapshot. It is a living system that gets better every time someone uses it seriously enough to notice where it fails.

---

## Limitations

---

- Single institution. Epic source system. Generalizability to non-Epic OMOP sites is untested.
  - The 0-of-9 fitness baseline has been partially re-measured (200+ projects validated through the blind reproduction protocol, with improving match rates) but a formal re-test of the original 9 projects has not been published.
  - The archaeological scan captures table-level gaps. Column-level and filter-level gaps emerge from the operational feedback loop, not from the scan.
  - The CDM row growth from 3.78B to 8.42B includes legitimate coverage expansion and also reflects the Clarity-to-Caboodle migration (described in the companion poster, “The Undiscovered Caboodle”), which increased rows through extraction completeness rather than gap closure.
  - The convergence claim is based on pattern discovery rate, not a formal coverage metric. We do not have a denominator for “all possible patterns.”
  - The AI agent’s quality is bounded by the knowledge base’s completeness. For domains with sparse pattern coverage, the agent produces less reliable SQL.
  - The build remains a single-analyst operation. The bus factor concern that motivated the architecture has not been fully resolved — the system is more resilient than one person’s memory, but it still depends on one person’s judgment for pattern triage and fix routing.
- 

## References

---

1. Observational Health Data Sciences and Informatics. The Book of OHDSI. Chapter 6: Extract Transform Load. 2024. [ohdsi.github.io/TheBookOfOhdsi/](https://ohdsi.github.io/TheBookOfOhdsi/)
  2. Kahn MG, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS. 2016;4(1):1244.
  3. Snowflake Inc. Snowflake Cortex Code. 2024. [docs.snowflake.com/en/user-guide/cortex-code/](https://docs.snowflake.com/en/user-guide/cortex-code/)
  4. Carlson R. OMOP VI: The Undiscovered Caboodle. Epic UserWeb. 2026. Thread 147477.
-