

The Undiscovered Caboodle

Validating a Parallel OMOP ETL Migration from Epic Clarity to Caboodle

Technical White Paper

September 2026

Roger Carlson

SASS Honest Broker Analytics

Corewell Health Research Institute, Grand Rapids, MI

Summary

We hypothesized that Epic Caboodle is a materially better OMOP CDM source than Epic Clarity for most clinical domains. To test this, we implemented dual ETL pipelines within a single dbt project, running both Clarity and Caboodle extractions against an identical 15,162-patient diverse cohort at Corewell Health (22 hospitals, two merged Epic instances). We compared 14 OMOP CDM v5.4 domains across row count parity, record-level concordance, concept mapping completeness, patient coverage, and visit detail linkage.

Results confirmed the hypothesis. Caboodle produced more clinical rows in 10 of 14 domains, with record-level concordance exceeding 97% for the three highest-volume domains (Measurement, Visit, Observation). The differences are architectural: Caboodle consolidates multiple extraction pathways per fact table, includes modules for ADT-level visit detail and population health SDOH, and pre-computes linkages that require complex multi-table traversal in Clarity. Clarity retained superiority for procedures (charge-level CPT granularity) and clinical notes (16% more ambulatory/hospital notes). The production system now runs as a Caboodle-primary hybrid with Clarity supplement for specific domains, serving 4.3 million patients across 7.0 billion clinical rows.

Introduction

Epic Systems provides two analytical data platforms: Clarity and Caboodle. Clarity is a relational reporting database that exposes Epic's internal transactional tables in a queryable format. Caboodle is a star-schema analytics warehouse built by Epic's data warehousing team, designed specifically for reporting and analytics consumption.

The conventional wisdom among OMOP CDM implementers at Epic sites holds that Clarity is the "gold standard" source for ETL development. This belief rests on three assumptions: (1) Clarity contains everything Caboodle contains because Caboodle is derived from Clarity, (2) Clarity provides finer granularity because it exposes the transactional layer, and (3) established community ETL conventions (including AllofUs submissions) use Clarity.

We challenged all three assumptions by building parallel ETL pipelines and comparing their outputs empirically.

Why Caboodle Can Be Richer Than Clarity

The premise that Caboodle is merely a “derived copy” of Clarity is architecturally incorrect. While Caboodle draws source data from the same Epic operational database as Clarity, it is a separately engineered star-schema model built by Epic’s analytics team. Four structural differences create genuine data advantages:

1. **Module inclusion.** Caboodle contains analytic modules that have no Clarity equivalent. PopulationHealthSDOH (social determinants of health from validated screening instruments) and ADT-level visit detail (Admission-Discharge-Transfer events) exist only in Caboodle’s dimensional model.
2. **Pre-computed linkage.** Caboodle’s star schema pre-joins data that requires complex multi-table traversal in Clarity. Hospital encounter events (conditions, measurements, drug administrations) link directly to ADT-level visit detail through foreign keys in Caboodle’s fact tables. In Clarity, reproducing this linkage requires navigating `PAT_ENC_HSP → PAT_ENC_HSP_2 → ED_ICD_VISIT → HSP_ACCT_MULT_DRGS` and similar chains.
3. **Supplementary data sources.** Caboodle receives data from pathways that bypass Clarity entirely. Patient-reported outcomes, population health imports, and certain clinical documentation flows write to Caboodle dimensions without corresponding Clarity tables.
4. **Extraction completeness.** Each Caboodle fact table is populated by multiple extraction packages that draw from different Clarity table hierarchies. An analyst writing Clarity SQL discovers one path. Epic’s extraction engineers, working over years, found all of them. A Caboodle fact table consolidates every path its extraction packages identified. We demonstrate this mechanism concretely in Section 3.5.1, where it explains the largest single disparity in our comparison.

The relevant question: what does Epic’s analytics engineering team add to the raw transactional data when building their star schema? The answer: structure, linkage, supplementary modules, and extraction completeness.

Methods

Study Design

We implemented a dual-build framework within a single dbt (data build tool) project at Corewell Health. Both pipelines read from production bronze sources (`CARE_BRONZE_CLARITY_PROD`, `CARE_BRONZE_CABOODLE_PROD`) and write to a shared target database using suffixed tables (`_CLA`, `_CAB`). A shared patient cohort ensured identical populations in both pipelines.

Patient Cohort

The development cohort comprised 14,001 adult patients enrolled in the All of Us Research Program at Corewell Health, supplemented with 1,161 of their offspring identified through maternal-fetal linkage. The baby supplement ensured coverage of pediatric and neonatal clinical domains (birth episodes, immunizations, NICU measurements) that an adult-only cohort would miss.

This cohort was selected because: (1) it spans the full age range from neonates to centenarians (birth years 1917-2026), (2) the adult component has been externally validated through All of Us data quality submissions, and (3) it is small enough to enable rapid iteration (sub-10-minute build times for both pipelines).

Comparison Framework

We compared all 14 OMOP CDM v5.4 clinical domains across five dimensions:

1. **Row count parity.** Total rows per domain in each pipeline, with percentage difference.
2. **Record-level concordance.** Using distinct (person, date, concept) triples, we measured what proportion of clinical events in one pipeline also appeared in the other.
3. **Concept mapping completeness.** For each domain, the percentage of rows with a non-zero standard concept_id (indicating successful vocabulary mapping).
4. **Patient coverage.** Whether the same patients appear in both pipelines, and whether one system captures patients the other misses.
5. **Visit detail linkage.** The percentage of hospital clinical events linked to ADT-level visit detail records.

ETL Specification

Both pipelines use identical vocabulary mapping logic (Source-to-Concept Maps, OHDSI standard vocabularies v5.0, 20250101 release). The transformation logic follows OMOP CDM v5.4 conventions with institution-specific source-to-concept mappings maintained in a shared STCM table (6,960 mappings at time of comparison).

Importantly, this comparison is restricted to clinical EHR data. All records in both pipelines carry type_concept_id 32817 (EHR). Neither pipeline draws from claims or billing transaction feeds. Caboodle does contain claims-related tables (BillingTransactionFact, ClaimFact) that have no Clarity equivalent; incorporating these would increase the measured difference. The results presented here represent a conservative estimate of Caboodle's data richness relative to Clarity.

The Clarity pipeline draws from 47 source tables across encounter, diagnosis, procedure, medication, measurement, note, device, and specimen domains. The Caboodle pipeline draws from 38 fact and dimension tables covering the same clinical territory.

Both pipelines are available for inspection on Epic UserWeb (Thread 147477), including compiled SQL and dbt model definitions.

AI-Augmented Development

The Caboodle pipeline was built by a single analyst using Snowflake Cortex Code (an AI-augmented development environment). The AI assistant generated initial dbt model scaffolds from source table documentation, identified join relationships from Caboodle’s foreign key metadata, and ran iterative validation queries comparing outputs between pipelines. All AI-generated code was reviewed, tested against the comparison framework, and validated against the established Clarity pipeline before acceptance. The full development cycle (scaffold, implement, compare, validate) completed in under 3 weeks.

Production Deployment

Based on the comparison results, we deployed a hybrid production architecture: Caboodle-primary for all domains except PROCEDURE_OCCURRENCE (where Clarity’s charge-level CPT granularity produces 3.7x more events) and NOTE (where Clarity captures 16% more ambulatory and hospital notes). No domain draws from both sources simultaneously; de-duplication is structural rather than algorithmic.

Results

Domain-Level Row Count Comparison

Table 1. Row Count Comparison: Clarity vs. Caboodle (15,162-patient cohort, April 2026)

Domain	Clarity Rows	Caboodle Rows	Difference	Verdict
DEATH	434	434	0%	Exact match
PERSON	15,153	15,162	+0.1%	Near-identical
VISIT_OCCURRENCE	2,496,276	2,517,620	+0.9%	Near-identical
MEASUREMENT	34,647,555	34,982,517	+1.0%	Near-identical
DRUG_EXPOSURE	1,920,575	2,038,493	+6.1%	CAB richer
SPECIMEN	56,269	49,891	-11.3%	CLA richer
CONDITION_OCCURRENCE	7,054,089	7,840,129	+11.1%	CAB richer (problem lists)
NOTE	5,583,731	4,672,764	-16.3%	CLA richer
OBSERVATION	4,860,177	5,703,797	+17.4%	CAB richer (ICD routing)
PROVIDER	271,738	320,842	+18.1%	CAB is superset
DEVICE_EXPOSURE	158,056	274,410	+73.6%*	CAB structured LDA + surgical supplies (Section 3.5.1)
PROCEDURE_OCCURRENCE	1,125,598	302,147	-73.2%	CLA charge-level
CARE_SITE	1,103	3,981	+260.9%	Different ID spaces
VISIT_DETAIL	280,545	3,365,979	+1,099.8%	Non-equivalent models

Caboodle produced more rows in 10 of 14 domains. Clarity was superior in 3 domains (PROCEDURE_OCCURRENCE, NOTE, SPECIMEN). One domain (DEATH) was an exact match.

*The DEVICE_EXPOSURE comparison captures only the LDA (Lines/Drains/Airways) and non-LDA modules available at comparison time. Post-comparison, we deployed surgical device pipelines from Caboodle's SurgicalSupplyUseFact, expanding production DEVICE_EXPOSURE to over 2.5M total rows (see Section 3.5.1).

Record-Level Concordance

To assess whether the row count differences represent genuinely additional clinical events versus duplicates or coding artifacts, we computed record-level concordance using distinct (person_id, date, concept_id) triples.

Table 2. Record-Level Concordance by Domain

Domain	CLA Distinct Events	CAB Distinct Events	Overlap	% CLA Found in CAB	% CAB Found in CLA
VISIT_OCCURRENCE	2,252,752	2,264,546	2,248,919	99.8%	99.3%
OBSERVATION	3,620,214	3,626,028	3,611,907	99.8%	99.6%
MEASUREMENT	10,699,571	10,736,620	10,468,390	97.8%	97.5%
CONDITION_OCCURRENCE	5,077,492	5,270,166	4,905,852	96.6%	93.1%
DRUG_EXPOSURE	1,438,645	1,591,663	1,355,703	94.2%	85.2%
NOTE	2,412,451	2,315,450	2,135,090	88.5%	92.2%

For the three highest-volume domains (MEASUREMENT, VISIT_OCCURRENCE, OBSERVATION), concordance exceeds 97%. The clinical events are substantively the same. Additional rows in Caboodle represent genuinely new data (problem list entries for CONDITION, additional medication sources for DRUG, ICD-routed observations) rather than duplicates.

Concept Mapping Completeness

Table 3. Concept Mapping Rates (Primary Concept Field)

Domain	Clarity Mapping Rate	Caboodle Mapping Rate
CONDITION_OCCURRENCE	100.00%	100.00%
PROCEDURE_OCCURRENCE	100.00%	100.00%
DRUG_EXPOSURE	100.00%	100.00%
OBSERVATION	100.00%	100.00%
NOTE (note_type)	100.00%	100.00%
MEASUREMENT	99.95%	99.94%
DEVICE_EXPOSURE	99.95%	99.97%
NOTE (note_class)	97.35%	99.87%

“100% mapping” means zero rows with concept_id = 0. This is achieved because the ETL excludes source records that cannot be mapped to a standard vocabulary concept. All source values are retained in the source_value fields; unmapped records are excluded from the CDM table rather than stored with concept_id = 0. This is a design choice: we prioritize CDM fidelity over source completeness, with the understanding that researchers requiring unmapped data can query the source systems directly via our semantic view layer (ACQUIRE).

Patient Coverage

Of 15,162 patients in the Caboodle cohort, all 15,153 Clarity patients were a perfect subset. The 9 additional Caboodle-only patients were all neonates (born 2018-2023) captured through Caboodle's PatientDim via maternal-fetal linkage pathways not present in the Clarity ETL's patient extraction logic.

This finding is notable because the same patient driver was provided to both pipelines. Caboodle's dimensional model resolved 9 additional neonatal patients from the maternal linkage that Clarity's transactional patient table did not surface. Both pipelines received the same input; Caboodle produced more complete output. This extends the hypothesis beyond clinical event data into patient identification itself: Caboodle's dimensional model surfaces patients that can be easily missed in Clarity's relational model.

Device Exposure Detail

At comparison time, the 73.6% row count advantage for Caboodle in DEVICE_EXPOSURE reflected two device categories. Post-comparison deployment of surgical device pipelines (Section 3.5.1) expanded this to three:

Device Category	Clarity	Caboodle	Unique Device ID Coverage
LDA (Lines, Drains, Airways)	Limited	129,742	99.95%
Non-LDA (radiopharmaceuticals, blood products, contrast)	158,056	144,668	0%
Surgical Devices (instruments + implants, post-comparison)	920,902*	2,225,095	0%

*Clarity surgical device count reflects V_LOG_SUPPLIES_IMPLANTS (a reporting view covering one of three extraction pathways). Direct extraction package queries yield higher counts but remain below Caboodle's case coverage (see Section 3.5.1).

LDA devices (peripheral IVs, endotracheal tubes, enteral tubes, supraglottic airways) are tracked through Caboodle's OutpatientDeviceFact with individual device placement identifiers. Clarity contains equivalent structured data in IP_LDA_NOADDSINGLE (10.3M rows, 1.7M patients) — a table we did not discover until post-comparison investigation. This is itself an instance of the extraction completeness pattern: the data exists in Clarity, but finding it requires knowing which table to query. Non-LDA devices (administered as part of procedures or charges) exist in both systems without individual serial numbers.

3.5.1 Surgical Supply Data: The Extraction Completeness Mechanism

The 73.6% Caboodle advantage in DEVICE_EXPOSURE was the largest unexplained disparity in our comparison. Three months after the comparison was frozen, the OHDSI Surgical & Perioperative Working Group (Evan Minty, lead) proposed using device tables to differentiate major from minor surgical procedures — a clinical question that required exactly the kind of surgical supply data Caboodle contained but our Clarity pipeline did not. This gave us both a direction and a clinical motivation to investigate the disparity.

The investigation revealed why our Clarity code was missing data. Our Clarity DEVICE_EXPOSURE pipeline queried V_LOG_SUPPLIES_IMPLANTS — a reporting view that covers one extraction path. Caboodle’s SurgicalSupplyUseFact is populated by three separate extraction packages, each drawing from a different Clarity table hierarchy: one traversing OR_LOG_LN_IMPLANT → OR_LNLG_IMPLANTS → OR_IMP (inventory-tracked implants), one for ad-hoc implants without inventory records, and a third pulling from OR_PKLST → OR_PKLST_SUP_LIST (pick-list supplies). Our Clarity query drew from one of these three paths. The other two paths — representing different classes of surgical supply documentation — were invisible to us until we inspected Epic’s extraction metadata in EpicDW.

This pattern — an analyst discovering one path while Caboodle consolidates all of them — turned out to be the central finding of the entire study. The DEVICE_EXPOSURE disparity was the most visible instance, but the same mechanism operates wherever a Caboodle fact table sits on top of multiple extraction packages.

Caboodle’s surgical supply tracking system (SurgicalSupplyUseFact + SurgicalSupplyDim) contains 32,028,403 use records across 444,155 distinct supply items spanning 1.1 million surgical cases. This data qualifies as OMOP Device domain content per CDM v5.4’s definition: “a foreign physical object or instrument used for diagnostic or therapeutic purposes.”

We profiled one category — Endomechanical Devices (staplers, trocars, energy devices, snares) — to assess the clinical utility that motivated the investigation:

Functional Group	Distinct Items	Surgical Cases
Energy Devices (LigaSure, Harmonic, Hot Shears)	643	142,952
Clip Appliers	1,040	139,626
Forceps/Graspers	322	134,610
Staplers (GIA, EEA, Echelon)	2,228	130,783
Trocars/Ports	448	114,740
Snares (polypectomy)	116	77,648
Total Endomechanical	8,596	420,882

To test whether device signatures can differentiate surgical procedure complexity — a hypothesis proposed by the OHDSI Surgical & Perioperative Working Group — we compared device presence across three procedure types that share overlapping ICD-10-PCS/SNOMED mappings:

Procedure Type	N Cases	Stapler%	Energy%	Trocar%	Snare%
Colectomy (resection)	7,684	97.7%	93.1%	58.5%	1.0%
Cholecystectomy	26,746	5.4%	25.1%	71.9%	0.1%
Colonoscopy (diagnostic/polypectomy)	34,376	0.3%	0.9%	0.1%	95.2%

The signal is unambiguous: stapler presence identifies resection cases at 97.7% sensitivity and 99.7% specificity relative to diagnostic colonoscopy. This confirms that surgical supply data can serve as an empirical classifier for procedure complexity where ICD-10-PCS procedure codes produce clinically incoherent groupings.

Comparison with Clarity’s nearest equivalent (V_LOG_SUPPLIES_IMPLANTS view, 27.1M total rows across all supply types) shows Caboodle provides 1.9x the row volume and 10x the item catalog granularity (8,596 vs. ~845 distinct endomechanical items). Caboodle’s SurgicalSupplyDim provides pre-classified TYPE and IMPLANT flags that enable automated pipeline routing without manual item-by-item classification. Clarity’s flat item listing (OR_IMP.IMPLANT_NAME) requires each implementer to independently determine which items are instruments, which are implants, and which are consumables.

Post-comparison, we built and deployed both surgical device pipelines using STCM-driven concept mapping:

Pipeline	STCM Mappings	Production Rows	Patients	SNOMED Concepts
Instruments (endomechanical)	183	1,857,884	296,155	18
Implants (typed categories)	4,014	367,211	158,407	25
Total surgical devices	4,197	2,225,095	—	43

Comparing these Caboodle production results with Clarity source coverage reveals the extraction completeness mechanism directly. For implants, the best Clarity pathway (OR_LOG_LN_IMPLANT → OR_LNLG_IMPLANTS → OR_IMP, matching the actual extraction package logic) covers 77,927 surgical cases with 121,935 item rows after deduplication. Caboodle covers 70,551 cases with 119,551 rows — near-parity (1.10x case difference, 1.02x item difference). For instruments, the pattern reverses slightly: Clarity produces 2.14M rows across 407K cases while Caboodle produces 2.26M rows across 403K cases (Caboodle 5% richer due to field preparation recording). Our original Clarity PULL query (against V_LOG_SUPPLIES_IMPLANTS, a reporting view) captured only 77,528 cases — 19% of what was available in Clarity’s own source tables. The reporting view aggregates away the transaction-level detail that the extraction packages preserve.

This surgical supply data expands production DEVICE_EXPOSURE from 112,639 rows (LDA-only) to over 2.2 million — representing the single largest clinical data source we identified in Caboodle during this study.

Production Deployment Metrics

Following the comparison, we deployed the Caboodle-primary hybrid architecture to the full patient population.

Table 4. Production CDM Metrics (July 2026)

Metric	Value
Total patients	4,277,374
Total clinical rows	6.98 billion
Hospital visit detail linkage	99.9%
Note class mapping rate	99.86%
LDA device unique ID coverage	99.95%
Curation checks passing	320 of 320 (0 errors)
Build time (full pipeline)	~90 minutes

Discussion

The Derived-System Paradox

Reviewer feedback on this study raised a foundational question: if Caboodle is derived from Clarity, how can it be superior? The answer is that “derived from” does not mean “subset of.” Epic’s data warehousing team performs substantial engineering work when building Caboodle from operational data: star-schema dimensional models, analytic modules with no Clarity equivalent, pre-computed temporal linkages, and — most significantly — extraction completeness.

The extraction completeness mechanism, demonstrated concretely in Section 3.5.1, is the key to understanding the entire study. The objection assumes that a Clarity query and a Caboodle fact table draw from the same source. They don’t. A Caboodle fact table draws from every source that Epic’s extraction engineers identified over years of development. A Clarity query draws from whichever tables the analyst discovered. The comparison is analyst-discovery vs. vendor-engineering. One is bounded by a single person’s time and documentation access. The other is bounded by a team’s accumulated institutional knowledge of the operational schema.

Our surgical device investigation (Section 3.5.1) makes this concrete. Our Clarity PULL query used `V_LOG_SUPPLIES_IMPLANTS` — a reporting view that covers one of three extraction pathways feeding `SurgicalSupplyUseFact`. That view captured 77,528 surgical cases. When we reconstructed the full extraction logic directly from Clarity’s source tables (`OR_LOG`, `OR_LOG_LN_IMPLANT`, `OR_LNLG_IMPLANTS`, `OR_IMP`, `OR_PKLST`, `OR_PKLST_SUP_LIST`, `OR_SPLY`, and `PAT_OR_ADM_LINK` — 8 tables, plus ZC lookup joins), we recovered 407,198 cases. That is 5.3x what our original query found — and near-identical to Caboodle’s 402,993 cases.

Domain	Clarity (V_LOG view)	Clarity (full extraction)	Caboodle	Ratio (full vs. CAB)
Instrument cases	77,528	407,198	402,993	1.01x
Instrument rows	—	2,139,451	2,257,164	0.95x
Implant cases	—	77,927	70,551	1.10x
Implant rows	—	121,935	119,551	1.02x

When both systems are queried with equivalent extraction logic, they produce functionally identical clinical data. The 5% Caboodle instrument advantage represents surgical field preparation items (every supply opened or staged) versus Clarity’s charted-usage-only recording. The complexity difference is structural: the Caboodle path requires 3 fact/dimension tables with linear joins; the Clarity path requires 8 base tables with multi-path join logic and careful deduplication (we identified two independent fanout sources that inflated naive Clarity counts by 2.4x before correction).

This is the central finding: the data exists in Clarity, but practically speaking, an analyst will not find all of it. Our original Clarity query captured 19% of what Clarity’s own tables contained. We did not discover the missing 81% until we inspected Epic’s extraction metadata — the same metadata that Caboodle’s extraction packages already incorporate. Caboodle does not add data that Clarity lacks. It adds the engineering that makes all of Clarity’s data discoverable.

Domains Where Clarity Remains Superior

Three domains favor Clarity:

PROCEDURE_OCCURRENCE (-73.2%): Clarity captures procedure-level detail from charge records, producing one row per CPT code billed. Caboodle’s ProcedureFact captures surgical/procedural events at a higher level of abstraction. For research requiring charge-level granularity (e.g., counting specific CPT codes for cost analysis), Clarity remains necessary. However, the surgical supply data (Section 3.5.1) demonstrates that Caboodle compensates for this procedural abstraction by providing device-level detail that can differentiate procedure complexity empirically — a capability Clarity’s charge-level CPT codes cannot provide.

NOTE (-16.3%): Clarity’s HNO_INFO table captures ambulatory and hospital notes more completely than Caboodle’s DocumentFact. The gap is concentrated in ambulatory progress notes and emergency department documentation. Caboodle’s note routing handles inpatient documentation well but has gaps in outpatient settings.

SPECIMEN (-11.3%): A minor difference attributable to source table coverage in Caboodle’s SpecimenFact versus Clarity’s SPECIMEN/SPEC_HIST tables.

The Hybrid Architecture

Our production system draws each domain from exactly one source. This eliminates de-duplication concerns entirely: a CONDITION row comes from Caboodle or Clarity, never both. The routing is configured at the dbt model level (each domain model has a single source configuration). Switching a domain’s source requires changing one model file and rebuilding.

Generalizability

Three findings generalize to any Epic site running both Clarity and Caboodle:

1. **Caboodle and Clarity do not produce identical CDMs.** Sites should not assume a simple port will be equivalent. A parallel validation framework (same cohort, both pipelines, domain-level comparison) is the minimum necessary to characterize differences before switching.
2. **A hybrid architecture may be optimal.** Neither source is universally superior across all OMOP domains. The specific domains where Clarity excels (procedures, notes) may differ by site depending on local Epic configuration and clinical workflow documentation patterns.
3. **The parallel validation framework is reproducible.** Any Epic site running both systems can implement the same comparison methodology. The dbt project structure and compiled SQL are available on Epic UserWeb (Thread 147477) for direct adaptation.

Limitations

- This study was conducted at a single health system with a specific Epic configuration. Domain-level advantages may vary by site.
- The development cohort (15,162 patients) is a convenience sample from the All of Us program, which may not represent the full demographic distribution of the health system.
- The 3-week development timeline reflects AI-augmented development; sites without similar tooling should expect longer timelines.
- Record-level concordance uses (person, date, concept) triples, which does not capture value-level agreement for measurements or dose-level agreement for drugs.
- The study compares ETL outputs, not the underlying source data. Some row count differences likely reflect Clarity query incompleteness rather than intrinsic Caboodle superiority. As demonstrated by the surgical device analysis (Section 3.5.1), a single Caboodle fact table may consolidate extraction pathways that our Clarity queries missed entirely. The practical finding holds regardless: the data exists in Clarity only in principle. Accessing it requires reverse-engineering Epic's extraction metadata.

Conclusion

Epic Caboodle is a materially better OMOP CDM source than Epic Clarity for most clinical domains. It produces more clinical rows in 10 of 14 domains, provides ADT-level visit detail linkage (99.9%) and tracks LDA devices with individual identifiers (99.95% coverage). Note classifications map more completely (99.87% vs 97.35%). Record-level concordance exceeds 97% for the three highest-volume domains, confirming that the additional rows represent genuine clinical data rather than artifacts.

The advantage is architectural: Caboodle is a separately engineered analytics model that adds structure, linkage, supplementary modules, and extraction completeness to the transactional source data. Each Caboodle fact table consolidates multiple Clarity extraction pathways that we did not discover until inspecting Epic's extraction metadata. A hybrid architecture (Caboodle-primary with Clarity supplement for procedures and notes) captures the best of both systems.

Beyond the initial comparison, Caboodle's SurgicalSupplyUseFact contains 32 million surgical supply use records that qualify as OMOP DEVICE_EXPOSURE content. We deployed 2.2 million of these (filtered through 4,197 STCM concept mappings) to production. The deployed data is 1.9x richer than Clarity's equivalent, with 10x the item catalog granularity, and can empirically differentiate surgical procedure complexity — a capability the OHDSI community has identified as an unmet need. This finding reinforces the central thesis: Caboodle is architecturally richer than Clarity, with clinical data discovery opportunities that continue to emerge as the platform is systematically profiled.

AI Disclosure

This white paper was developed with AI assistance throughout. Snowflake Cortex Code (Claude Sonnet and Opus models via Snowflake's AI development environment) served as the primary development tool for the Caboodle ETL pipeline, comparison queries, and root cause analysis. A single analyst built the full parallel pipeline, ran all domain-level comparisons, and investigated every outlier using AI-assisted iterative SQL generation. The AI accelerated dbt model authoring, cross-source comparison queries, data profiling, and SNOMED vocabulary mapping.

The white paper text was drafted with AI writing assistance (Claude Opus via Cortex Code). All data, findings, and conclusions were verified by the author against live Snowflake query results. Verification SQL is provided in Appendix A.

The OHDSI poster and brief report associated with this work were submitted with full AI disclosure per OHDSI symposium guidelines. The Corewell Health Research Institute's prohibition on AI use in scientific publications does not apply to this document: it is a technical white paper, not a peer-reviewed journal article.

Reproducing This Work

This white paper is intended to be reproducible by any Epic site running both Clarity and Caboodle on Snowflake (or any platform supporting dbt).

What you need:

- Read access to your site's Clarity and Caboodle bronze databases
- A dbt project targeting Snowflake (or your analytics warehouse)
- OHDSI standard vocabularies (v5.0 or later)
- A patient cohort for the comparison (we used All of Us enrollees; any defined cohort works)

What is available:

- Complete dbt project code (Clarity and Caboodle pipelines): Epic UserWeb, Thread 147477
- Compiled target SQL for both pipelines: same thread
- Verification queries: Appendix A of this document
- STCM mappings for surgical device classification: available on request

Steps:

1. Configure both pipelines against your bronze sources
2. Define a shared patient cohort and load it as a PATIENT_DRIVER table
3. Build both pipelines against the same cohort
4. Run the comparison queries from Appendix A, substituting your database and schema names
5. The domain-level verdicts will differ by site. That is expected and is itself a finding.

For implementation guidance specific to Clarity or Caboodle pipeline setup, CoCo (Cortex Code) implementation skills are available in the dbt project repositories.

References

1. Observational Health Data Sciences and Informatics. The Book of OHDSI. Chapter 6: Extract Transform Load. 2024. Available from: <https://ohdsi.github.io/TheBookOfOhdsi/>
2. Carlson R. OMOP VI: The Undiscovered Caboodle. Epic UserWeb. 2026. Thread 147477.
3. OMOP Common Data Model v5.4 Specification. OHDSI. Available from: <https://ohdsi.github.io/CommonDataModel/cdm54.html>

Appendix A: Verification Queries

All numbers in this manuscript were verified on July 17, 2026, by executing queries directly against the frozen comparison tables in CARE_RES_OMOP_DEV2_WKSP.OMOP (dev cohort, last altered April 13-14, 2026) and CARE_RES_OMOP_GEN_WKSP.OMOP (production).

Table 1 Verification

```
-- Row counts from frozen comparison tables
SELECT COUNT(*) FROM CARE_RES_OMOP_DEV2_WKSP.OMOP.<DOMAIN>_RAW_CLA;
SELECT COUNT(*) FROM CARE_RES_OMOP_DEV2_WKSP.OMOP.<DOMAIN>_RAW_CAB;
```

Table 2 Concordance

```
-- Record-level concordance (example: MEASUREMENT)
WITH cla_dist AS (
  SELECT DISTINCT PERSON_ID, MEASUREMENT_DATE, MEASUREMENT_CONCEPT_ID
  FROM CARE_RES_OMOP_DEV2_WKSP.OMOP.MEASUREMENT_RAW_CLA
),
cab_dist AS (
  SELECT DISTINCT PERSON_ID, MEASUREMENT_DATE, MEASUREMENT_CONCEPT_ID
  FROM CARE_RES_OMOP_DEV2_WKSP.OMOP.MEASUREMENT_RAW_CAB
)
SELECT
  (SELECT COUNT(*) FROM cla_dist) AS CLA_DISTINCT,
  (SELECT COUNT(*) FROM cab_dist) AS CAB_DISTINCT,
  (SELECT COUNT(*) FROM cla_dist c
   INNER JOIN cab_dist b ON c.PERSON_ID = b.PERSON_ID
    AND c.MEASUREMENT_DATE = b.MEASUREMENT_DATE
    AND c.MEASUREMENT_CONCEPT_ID = b.MEASUREMENT_CONCEPT_ID) AS OVERLAP;
```

Table 3 Concept Mapping

```
-- Concept mapping rate (example: CONDITION)
SELECT COUNT(*) AS TOTAL,
  SUM(CASE WHEN CONDITION_CONCEPT_ID = 0 THEN 1 ELSE 0 END) AS UNMAPPED
FROM CARE_RES_OMOP_DEV2_WKSP.OMOP.CONDITION_OCCURRENCE_RAW_CAB;
```

Device Category Analysis

```
SELECT
  CASE WHEN DEVICE_SOURCE_VALUE LIKE '%LDA%'
  THEN 'LDA' ELSE 'Non-LDA' END AS CATEGORY,
  COUNT(*) AS TOTAL,
  SUM(CASE WHEN UNIQUE_DEVICE_ID IS NOT NULL AND UNIQUE_DEVICE_ID != ''
  THEN 1 ELSE 0 END) AS HAS_UDI
FROM CARE_RES_OMOP_DEV2_WKSP.OMOP.DEVICE_EXPOSURE_RAW_CAB
GROUP BY 1;
```

Appendix B: Data Availability

The complete dbt project code for both pipelines (Clarity and Caboodle) and the compiled target SQL are available on Epic UserWeb, Thread 147477. Four attachments provide full transparency:

1. `clarity-dbt` — Clarity pipeline dbt model definitions
2. `clarity-sql` — Compiled target SQL for Clarity pipeline
3. `caboodle-dbt` — Caboodle pipeline dbt model definitions
4. `caboodle-sql` — Compiled target SQL for Caboodle pipeline

The frozen comparison tables remain available in CARE_RES_OMOP_DEV2_WKSP.OMOP with _CLA and _CAB suffixes (last altered April 13-14, 2026). Production CDM is maintained in CARE_RES_OMOP_GEN_WKSP.OMOP.