

Data Link Odyssey: Methods for Linking Related Records in OMOP CDM v5.4

Eugene Paulenkovich¹, Melanie Philofsky¹

¹ EPAM Systems

OHDSI 2026 Symposium

Abstract

Clinical data rarely exists in isolation. Systolic and diastolic blood pressure readings belong together; laboratory panels share a common order; cancer staging records are meaningful only when linked to their qualifying diagnosis. Connecting related clinical events within the OMOP Common Data Model (CDM) v5.4 is a persistent challenge that directly affects the reproducibility of federated observational studies. Through a structured review of CDM v5.4 documentation, the OHDSI Oncology Extension, community forum discussions, and ATLAS cohort-definition guidance, we identified and evaluated five distinct record linkage mechanisms: (1) PK/FK linkage including EPISODE/EPISODE_EVENT, (2) FACT_RELATIONSHIP, (3) oncology event linkage fields (`_event_id` and `_event_field_concept_id`), (4) concept hierarchies via `concept_ancestor`, and (5) custom tables and fields. Each mechanism was assessed against four criteria: domain coverage, directionality and cardinality, support in ATLAS and community tools, and implementation complexity. We present a structured comparison of linkage mechanisms and a practical decision guide. We conclude that the absence of written community conventions forces inconsistent implementations that undermine multi-site reproducibility, and we urge the OHDSI community to establish and maintain documented linkage conventions with worked examples.

Keywords: OMOP CDM, record linkage, FACT_RELATIONSHIP, EPISODE_EVENT, concept hierarchy, observational research, data interoperability, federated studies

1. Background

Clinical data rarely exists in isolation. Systolic and diastolic blood pressure readings belong together; laboratory panels share a common order; cancer staging records are meaningful only when linked to their qualifying diagnosis. This is the challenge of record linkage - explicitly connecting related clinical events within a structured data model.

The OMOP Common Data Model (CDM) v5.4, maintained by the Observational Health Data Sciences and Informatics (OHDSI) community, provides a standardized schema for representing patient-level clinical data from heterogeneous sources.¹ While the OHDSI community has established conventions for some linkage patterns (e.g., blood pressure pairs, cancer episodes via the Oncology Extension⁴), many others remain underspecified or undocumented.

This gap has real analytic consequences. Without consistent conventions, different sites implement the same linkage patterns differently. A federated study querying linked blood pressure pairs may return accurate results at Site A, empty results at Site B (which stores the same data but uses a different mechanism), and inflated results at Site C (which conflates co-occurring but unrelated readings). Analysts must guess intent, creating ambiguous cohort logic and eroding confidence in results. This directly contributes to the reproducibility crisis in federated observational research.¹

Existing documentation for each mechanism is scattered across the CDM specification, the OHDSI Book of OHDSI, Oncology Extension guidance, ATLAS wiki pages, and community forum threads.^{1,2,4,5,6} No single resource consolidates the available strategies, compares them systematically, or provides implementers with actionable selection criteria. This paper addresses that gap.

2. Methods

2.1 Literature and Documentation Review

We conducted a structured review of sources documenting record linkage capabilities in OMOP CDM v5.4. The review corpus comprised: the OMOP CDM v5.4 specification¹; The Book of OHDSI²; the OHDSI Oncology Extension documentation⁴; ATLAS cohort definition guidance from the ATLAS GitHub wiki⁵; the OHDSI community forums, specifically the thread on EPISODE_EVENT adoption barriers (January 2026)⁶; and the OHDSI standardized vocabulary documentation.³ Sources were identified via direct review of official OHDSI documentation sites and supplemented by search of the OHDSI community forum using the terms 'record linkage', 'fact_relationship', 'episode_event', and 'event_id'.

2.2 Inclusion and Exclusion Criteria

A mechanism was included if it: (1) is available within the standard OMOP CDM v5.4 schema or an official OHDSI extension; (2) enables associating two or more clinical event records with one another; and (3) is documented in at least one official OHDSI source. Custom extensions were included as a category because they represent a common de facto practice that must be characterized for completeness. Mechanisms applicable only to demographic linkage (e.g., linking patients across systems) were excluded as outside scope.

2.3 Evaluation Framework

Each identified mechanism was evaluated against four criteria derived from the analytic requirements of federated observational research:

1. Domain Coverage — which CDM clinical tables and domains the mechanism can operate across.
2. Directionality and Cardinality — whether the link is unidirectional or bidirectional, and whether it supports one-to-one, one-to-many, many-to-one, or many-to-many relationships.
3. Community Tool Support — whether the mechanism is natively accessible in ATLAS and other standard OHDSI community tools, without requiring custom database queries.
4. Implementation Complexity — the cognitive and engineering overhead required to correctly implement and query the mechanism, including join complexity and vocabulary knowledge requirements.

3. Results

3.1 Overview of Identified Mechanisms

Five distinct linkage mechanisms were identified within OMOP CDM v5.4 and its official extensions. Table 1 provides a structured comparison across all four evaluation criteria. The following subsections describe each mechanism in detail.

Table 1. Comparison of Record Linkage Mechanisms in OMOP CDM v5.4

Mechanism	Domain Coverage	Directionality / Cardinality	ATLAS / Tool Support	Complexity	Best Use Case
PK/FK Linkage (incl. EPISODE/ EPISODE_EVENT)	All clinical tables (universal)	Unidirectional (child → parent) Many-to-one	Full (visit context is core ATLAS feature)	Low	Grouping events within a visit; structured clinical courses (episodes)
FACT_RELATIONSHIP	Any domain pair (cross-domain)	Bidirectional by convention (2 rows per link) Many-to-many	None — not exposed in standard tools	High (multi-table joins, narrow concept catalog)	Typed, explicit, cross-domain links (e.g., specimen → measurement)
Oncology Event Linkage Fields (_event_id/ _event_field_ concept_id)	MEASUREMENT and OBSERVATION only	Unidirectional (child → parent only) Many-to-one	None — not in standard ATLAS	Medium (field_concept_id routing logic)	Measurement / Observation modifiers, oncology attributes, tumor staging
Concept Hierarchies (concept_ancestor)	Any vocabulary-backed domain	Bidirectional (ancestor ↔ descendant) One-to-many	Full — core ATLAS "Include Descendants" feature	Low–Medium (vocabulary knowledge required)	Cohort grouping, lab panel roll-up, drug class membership

Mechanism	Domain Coverage	Directionality / Cardinality	ATLAS / Tool Support	Complexity	Best Use Case
Custom Tables and Fields	Unlimited (site-defined)	Site-defined (any pattern)	None — invisible to all standard tools	Low to use; High to maintain and document	Last resort only; site-specific needs not met by above

3.2 Mechanism 1: PK/FK Linkage (including EPISODE and EPISODE_EVENT)

The primary key / foreign key (PK/FK) relationship represents the most fundamental linkage pattern in OMOP CDM. Every clinical event table (CONDITION_OCCURRENCE, MEASUREMENT, DRUG_EXPOSURE, PROCEDURE_OCCURRENCE, OBSERVATION, and others) carries a `visit_occurrence_id` foreign key that connects each event to its parent VISIT_OCCURRENCE record.¹ This pattern is universal, universally supported in ATLAS, and requires only a standard JOIN on an indexed column.

CDM v5.4 extended PK/FK linkage with two new tables: EPISODE and EPISODE_EVENT. EPISODE represents a clinically meaningful course (e.g., a chemotherapy regimen, a pregnancy), while EPISODE_EVENT links individual clinical events into that course. EPISODE_EVENT is particularly valuable for oncology use cases where a treatment regimen episode is composed of multiple drug exposure events.

A critical adoption concern: EPISODE_EVENT does not use a direct foreign key to the linked event record. Instead, it uses an `episode_object_concept_id` (a `field_concept_id`) to indicate which table the `event_id` refers to.⁶ This indirect indirection pattern requires implementers to perform an extra lookup step before resolving the link, and has been identified as a significant barrier to adoption in the OHDSI community forums.⁶

Strengths: universal domain coverage; full ATLAS support; low implementation complexity; highly performant via indexed FKs. Limitations: `visit_occurrence_id` captures only the parent visit, not the semantic relationship between two sibling events (e.g., which specific measurement pairs belong together). EPISODE_EVENT's `field_concept_id` indirection adds cognitive and query overhead.

3.3 Mechanism 2: FACT_RELATIONSHIP

FACT_RELATIONSHIP is the only native OMOP CDM mechanism that supports typed, explicit, cross-domain links between any two clinical facts. Each row asserts: "Record A in Domain 1 has relationship R with Record B in Domain 2." Bidirectionality is achieved by convention: two rows are inserted per link (A→B and B→A), enabling navigation from either side.¹

FACT_RELATIONSHIP is uniquely capable in two dimensions: it supports any domain pair (e.g., SPECIMEN to MEASUREMENT, DRUG_EXPOSURE to CONDITION_OCCURRENCE) and it encodes the relationship type via a `relationship_concept_id`, enabling semantically meaningful differentiation between link types (e.g., 'derived from' vs. 'has indication').

Despite this theoretical power, FACT_RELATIONSHIP suffers from two significant practical limitations. First, it has no support in ATLAS or other standard community tools — relationships stored in FACT_RELATIONSHIP are entirely invisible to the cohort builder and most analytical pipelines. Second, its relationship concept catalog is narrow: of approximately 209 available relationship concepts, roughly 185 originate from SNOMED and represent person-to-person

relationships (parent-of, sibling-of, etc.), leaving only approximately 24 concepts applicable to event-level data linkage.³ The Appendix of this paper proposes an initial catalog expansion.

Strengths: only mechanism supporting typed cross-domain links; many-to-many capable; bidirectional by convention. Limitations: no ATLAS support; narrow concept catalog; high query complexity (multi-table joins required); no `person_id` on the table itself (requires anchor join to filter by patient).

3.4 Mechanism 3: Oncology Event Linkage Fields

Introduced in the OHDSI Oncology Extension⁴, MEASUREMENT and OBSERVATION tables in CDM v5.4 carry two event linkage fields: `measurement_event_id/observation_event_id` (the PK of the linked parent record) and `meas_event_field_concept_id/obs_event_field_concept_id` (a `field_concept_id` identifying which table the `event_id` refers to). Together, these fields create a direct many-to-one pointer: multiple modifier records can reference a single parent event.

The primary use case is oncology: multiple tumor attribute observations (T stage, N stage, M stage, grade) point back to a single EPISODE or CONDITION_OCCURRENCE record representing the primary cancer diagnosis. The pattern is also applicable to laboratory panels (measurement components pointing to a procedure occurrence) and any scenario where a Measurement or Observation record contextualizes or qualifies a parent event.

Limitations: the mechanism is restricted to MEASUREMENT and OBSERVATION as the child domain — records in other tables cannot carry `event_id` pointers. Links are strictly unidirectional (visible only from the child side). Like EPISODE_EVENT, `field_concept_id` routing requires implementers to determine the target table before resolving the join. The fields are not supported in standard ATLAS.

Strengths: direct, compact pointer with low storage overhead; supports many modifiers per parent event; consistent with Oncology Extension guidance. Limitations: unidirectional; domain-limited (MEASUREMENT/OBSERVATION child only); encodes no relationship type; no ATLAS support.

3.5 Mechanism 4: Concept Hierarchies via `concept_ancestor`

The CONCEPT_ANCESTOR table stores all ancestor-descendant concept pairs from OMOP standardized vocabularies (SNOMED, LOINC, RxNorm, and others), with minimum and maximum levels of separation.³ This mechanism enables grouping of clinical event records under a shared parent concept, supporting use cases such as: laboratory panel component grouping (all CBC components under the CBC panel LOINC concept), drug class membership (all opioids under the RxNorm opioid class), oxygen saturation readings across multiple device concepts, and questionnaire section grouping.

Concept hierarchies are the most ATLAS-native mechanism: the "Include Descendants" option in ATLAS cohort definitions directly leverages `concept_ancestor` to expand a concept set to include all descendants.²⁵ This makes concept hierarchies the recommended approach for cohort grouping tasks that do not require explicit event-to-event record linkage.

A critical distinction: `concept_ancestor` links vocabulary CONCEPTS, not individual event records. If the goal is to assert "these two specific measurement records came from the same laboratory order," concept hierarchies cannot provide that — they can only confirm that both

records share a common vocabulary ancestor. For explicit record-level linkage, Mechanisms 1 or 2 are required.

Strengths: full ATLAS support; bidirectional hierarchy traversal; no additional schema required; covers all vocabulary-backed domains. Limitations: links concepts, not individual records; requires vocabulary expertise; may not capture all relevant concepts if source mappings vary across sites.

3.6 Mechanism 5: Custom Tables and Fields

When no standard mechanism meets the analytic requirement, sites extend the CDM with custom tables or columns. This provides unlimited flexibility: any relationship type, any cardinality, any metadata can be represented. Common examples include adding a `bp_pair_id` column to `MEASUREMENT` to group blood pressure pairs, creating a separate `event_link` table for arbitrary relationship tracking, or adding order identifiers that link laboratory results back to their source requisition.

Custom extensions carry substantial costs. They are completely invisible to ATLAS and all standard community tools. They accumulate technical debt across CDM upgrade cycles. They require explicit documentation (including source mappings, population rules, and deprecation plans) to remain usable across teams and over time. Most critically for federated research: custom extensions are opaque to federated query runners — the linked data exists at the site but cannot be accessed by any standard federated query.

Strengths: maximum flexibility for site-specific requirements; no schema constraints. Limitations: no community tool support; invisible to federated queries; high documentation overhead; creates upgrade maintenance burden. Recommendation: use only as a last resort, after evaluating all standard mechanisms. Maintain a site extension registry and plan for eventual migration to a standard pattern.

4. Discussion

4.1 Practical Guide: Selecting the Right Mechanism

The five mechanisms are not interchangeable — each is suited to a distinct class of problems. The following decision guide summarizes the selection logic.

Table 2. Practical Selection Guide for Record Linkage Mechanisms

Clinical Scenario	Recommended Mechanism	Primary CDM Construct
Grouping clinical events within a visit	PK/FK — <code>visit_occurrence_id</code>	<code>VISIT_OCCURRENCE</code> (<code>visit_occurrence_id</code>)
Structuring a clinical course (treatment regimen, pregnancy)	<code>EPISODE</code> / <code>EPISODE_EVENT</code>	<code>EPISODE</code> , <code>EPISODE_EVENT</code> (<code>episode_object_concept_id</code>)
Typed, explicit cross-domain link (e.g., specimen → result)	<code>FACT_RELATIONSHIP</code>	<code>FACT_RELATIONSHIP</code> (<code>relationship_concept_id</code>)
Measurement or Observation modifier / qualifier	Oncology event fields (<code>_event_id</code>)	<code>MEASUREMENT</code> / <code>OBSERVATION</code> (<code>_event_id</code> , <code>_event_field_concept_id</code>)

Clinical Scenario	Recommended Mechanism	Primary CDM Construct
Tumor attributes, staging, grading	Oncology event fields (EPISODE anchor per Oncology Extension)	EPISODE, MEASUREMENT / OBSERVATION
Cohort grouping by drug class, lab panel, or diagnosis category	Concept hierarchy (concept_ancestor)	CONCEPT_ANCESTOR (ancestor_concept_id)
Site-specific need not met by any standard mechanism	Custom extension (last resort — document thoroughly)	Site-defined tables or columns

4.3 Limitations of This Review

The relationship concept catalog statistics cited (185 of 209 concepts being SNOMED person-to-person relationships) reflect a point-in-time review of the OHDSI Athena vocabulary browser and may change as the vocabulary evolves.³

4.4 Community Call to Action

The OHDSI community should prioritize establishing clear, documented conventions with worked examples for record linkage. Specifically, we recommend the following initiatives:

1. Publish a linkage conventions document as a THEMIS convention (or equivalent) — specifying the canonical mechanism for each common clinical linkage pattern (blood pressure pairs, lab panels, cancer staging, etc.).
2. Expand the FACT_RELATIONSHIP concept catalog — the CDM working group and vocabulary team should collaborate to define and publish an initial set of event-level relationship concepts (see Appendix for a proposed starting set).
3. Add ATLAS support for FACT_RELATIONSHIP — at minimum, a cohort builder filter enabling 'events linked to X via FACT_RELATIONSHIP' would dramatically increase adoption of this mechanism.
4. Create a CDM conformance test for linkage — the CDM release checklist should include validation queries confirming that key linkage patterns are consistently populated at conformance-tested sites.
5. Deprecate or document the EPISODE_EVENT field_concept_id indirection pattern — the CDM working group should either convert EPISODE_EVENT to use a direct FK or publish explicit documentation acknowledging and justifying the current design.

5. Conclusions

OMOP CDM v5.4 offers a range of mechanisms for linking related clinical events, each suited to a different class of problem. As a practical guide:

- Use PK/FK linkage (visit_occurrence_id) for grouping events within a visit — the universal, ATLAS-supported default.
- Use EPISODE / EPISODE_EVENT for structured clinical courses such as treatment regimens or pregnancies.
- Use FACT_RELATIONSHIP for typed, explicit, cross-domain relationships — accepting the absence of ATLAS support and the current concept catalog limitations.
- Use oncology event linkage fields (_event_id / _event_field_concept_id) for Measurement or Observation modifiers, especially in oncology workflows.

- Use concept hierarchies when the goal is cohort grouping in ATLAS — particularly for drug classes, diagnosis categories, or lab panels.
- Reserve custom extensions as a last resort, accepting mandatory documentation overhead and federated invisibility.

Critically, the absence of written community conventions means that different sites implement the same linkage patterns differently, directly contributing to analysis errors in federated research. No amount of individual site diligence can compensate for the absence of shared standards. The OHDSI community should treat linkage conventions with the same rigor as domain mapping conventions, and establish a standing working group to maintain and evolve these standards as the CDM grows.

This paper is offered as a starting point for that conversation. We invite the OHDSI community to collaborate on refining these recommendations into formal, versioned conventions that can be adopted across the network.

References

1. OHDSI Community. *OMOP Common Data Model v5.4*. OHDSI; 2021. <https://ohdsi.github.io/CommonDataModel/cdm54.html>
2. OHDSI Community. *The Book of OHDSI*. [Online; continuously updated]. <https://ohdsi.github.io/TheBookOfOhdsi/>
3. Reich C, Ostropolets A, Ryan P, et al. *OHDSI Standardized Vocabularies — a large-scale centralized reference ontology for international data harmonization*. J Am Med Inform Assoc. 2024;**31**(3):583–590. DOI: 10.1093/jamia/ocad247.
4. OHDSI. *Oncology Extension — OMOP Common Data Model*. [Online]. <https://ohdsi.github.io/CommonDataModel/oncology.html>
5. OHDSI. *ATLAS: Cohort Definitions*. GitHub Wiki. <https://github.com/OHDSI/Atlas/wiki/Cohort-Definitions>
6. Moinat M. *Episode Event*. CDM Builders — OHDSI Forums. January 9, 2026. <https://forums.ohdsi.org/t/episode-event/24824>

Appendix: Proposed Initial FACT_RELATIONSHIP Event-Level Relationship Concepts

The current FACT_RELATIONSHIP concept catalog contains approximately 24 concepts applicable to event-level data linkage. The following table proposes an initial set of additional relationship concepts to address the most commonly requested linkage patterns identified through community forum review. These are submitted for discussion and review by the OHDSI Vocabulary Working Group and CDM Working Group. Concept IDs marked 'Proposed' require formal assignment through the OHDSI vocabulary submission process.

Table 3. Proposed Initial Event-Level Relationship Concepts for FACT_RELATIONSHIP

Proposed Concept Name	Inverse Concept Name	Use Case	Example
Has paired measurement	Is paired measurement of	Link SBP and DBP from the same device event	MEASUREMENT (SBP) ↔ MEASUREMENT (DBP)
Derived from specimen	Has derived measurement	Link a lab result to its source specimen	MEASUREMENT ↔ SPECIMEN
Ordered via procedure	Has resulting measurement	Link a lab result to its order procedure	MEASUREMENT ↔ PROCEDURE_OCCURRENCE
Has indication	Is indication for	Link a drug exposure to the qualifying diagnosis	DRUG_EXPOSURE ↔ CONDITION_OCCURRENCE
Qualifies observation	Is qualified by observation	Link a modifier observation to its parent event	OBSERVATION ↔ (any domain)
Performed during visit	Has procedure performed	Explicit procedure-to-visit link beyond FK	PROCEDURE_OCCURRENCE ↔ VISIT_OCCURRENCE
Has device used	Was used in procedure	Link a device to the procedure it was used in	DEVICE_EXPOSURE ↔ PROCEDURE_OCCURRENCE
Sequential to	Precedes	Order sequential clinical events (e.g., pre/post measurements)	MEASUREMENT ↔ MEASUREMENT

Note: The OHDSI vocabulary submission process requires formal concept proposals through Athena and review by the vocabulary team. This appendix is intended as a discussion starting point, not a formal vocabulary submission. We encourage interested parties to engage with the OHDSI Vocabulary Working Group to advance these proposals.