Integrating clinical and laboratory research data using the OMOP CDM

Edward A. Frankenberger[1], Chun Yang[1], Vamsidhar Reddy Meda Venkata[1], Alyssa Goodson[1]
[1]Freenome

BACKGROUND

The use of real-world data (RWD) is expanding beyond its traditional roles in comparative effectiveness research and safety assessments[1]. Integrating biopharmaceutical laboratory research data with RWD in pre- and post-market settings is a relatively new use case that has the potential to improve clinical assay development[2,3,4], but its implementation is poorly described in current literature. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) has a well defined, generalized structure that is amenable to storing these data, along with a history of supporting biospecimen research[5,6]. In order to understand specimen research data in a clinical context during the process of assay development and inform improvements after release, we describe methodology by which healthcare, specimen, and research data are combined into a single OMOP CDM version 5.4 instance.

METHODS

Data Ingestion

Biospecimens are delivered along with sample metadata and clinical data for patients from whom the samples were collected. These data are separated and curated with different processes designed for each type of data. After curation, data are combined into a single analysis OMOP instance (figure 1). Follow-up data, when available, is ingested through their respective pipeline and combined with initial specimen collection data.

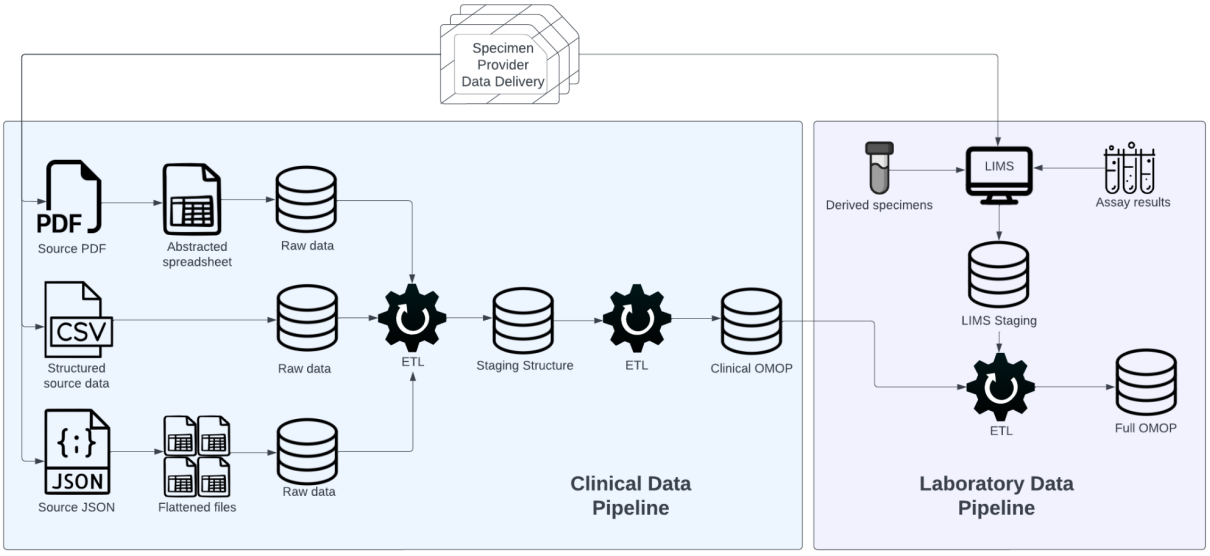

Figure 1. Overview of OMOP Extract, Transform, Load (ETL) process

Clinical data content and structure can vary widely. Data from commercial biobanking services may consist of a scanned PDF of a questionnaire completed prior to sample collection, generating a single point-in-time observation of subjective medical history, whereas samples from hospital systems may be

associated with codified, longitudinal healthcare information delivered as discrete data elements on a continual basis. Different methodologies based on source format were developed to modify clinical data in such a way that it can be ingested into a database, after which it undergoes an ETL process to produce a standardized, intermediate structure, followed by a second ETL to generate a standard OMOP database.

Data elements such as source patient ID, accession number, and volume can be expected from biospecimen providers in a relatively standardized format. These data are manually transferred by laboratory technicians into laboratory information management system (LIMS) software, which is used by clinical development teams to track assay results and the use of specimen material. Data from LIMS is regularly copied into a read-only database with a consistent schema, against which an ETL process was developed to move data into the specimen, measurement, and fact_relationship tables of an existing OMOP database.

Integrating Specimen Data

Specimens are often divided into subcomponents (e.g. separating plasma from whole blood) and further segregated into smaller aliquots used for testing. These procedures are documented in LIMS and made available to analysts in OMOP by modeling bidirectional parent/child relationships between a source specimen and its derived specimens in the fact_relationship table.

Collection events during which solid and/or liquid specimens and medical information are obtained from patients were modeled as encounters in the visit_occurrence table. To better establish relationships between specimens and clinical data, the specimen table was modified to include foreign key references to the visit_occurrence and visit_detail tables. Derived samples and aliquots inherited the visit_occurrence_id and (if available) visit_detail_id values from their parent samples.

Assay results and specimen metadata (e.g. freezer location, shipping conditions, etc.) were stored in the measurement table. A specimen record's primary key in OMOP was used as a foreign key reference in measurement.measurement_event_id to establish a relationship between records in the two tables, with measurement.meas_event_field_concept_id field set to 1147822 to indicate the table to which the key in measurement_event_id refers. Figure 2 provides a simplified representation of the relationships between clinical and laboratory data within the OMOP tables.

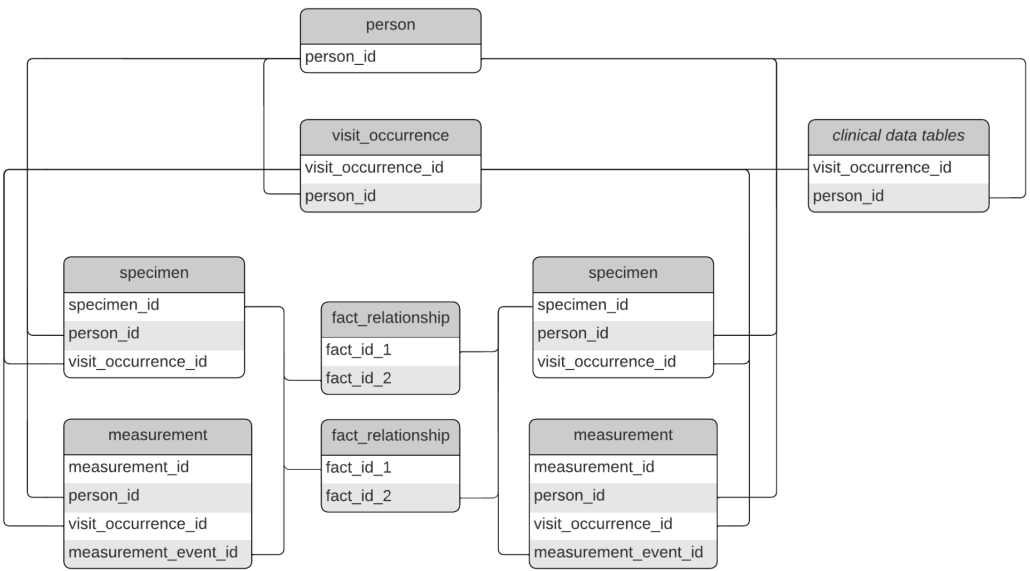

Figure 2. Simplified relationship diagram showing connections between clinical and laboratory data

Merging Data Sources

Referential integrity between LIMS and clinical data was successfully established through the development of surrogate keys composed of information available in both sources. A deterministic hash function used these surrogate keys, as described in table 1, as inputs to generate unique, consistent OMOP primary and foreign keys.

Table	Identifier	Surrogate Key	Sources
all	person_id	source patient ID + specimen provider name	Specimen provider
all	visit_occurrence_id	source patient ID + specimen provider name + date of collection	Specimen provider
specimen	specimen_id	specimen ID from LIMS	Specimen provider, local (derived sample)
fact_relationship	fact_id_[1-2]	specimen ID from LIMS	Specimen provider, local (derived sample)
measurement	measurment_id	source patient ID + specimen provider name + datetime + measurement + result	Clinical data, local (assay result)
measurement	measurement_event_id	specimen ID from LIMS	Specimen provider, local (derived sample)
observation	observation_id	source patient ID + specimen provider name + datetime + observation + result	Clinical data
observation	observation_event_id	specimen ID from LIMS	Specimen provider, local (derived sample)

Table 1. Surrogate key definitions and data sources

RESULTS

Codifying Data

A number of custom concepts were necessary to integrate laboratory research data into the OMOP CDM. New concepts describing organization-specific facts, such as assays under development, specialized assay results, and freezer names, were required. Additionally, parent to child and child to parent specimen relationship concepts (analogous to existing concepts 581436 and 581437) and metadata concepts for new fields in specimen were created.

Clinical development researchers rely on many patient survey responses, requiring specific semantic mapping standards. As compared to a typical OMOP instance in which patient-reported health histories

are codified to concepts in the Observation domain, we mapped diagnostic self-reports to concepts in the Condition domain to better align with user expectations. We relied heavily on type concept ID's to indicate which clinical data was sourced from a patient questionnaire (e.g. 32865) or the biospecimen provider (e.g. 32856) to preserve data provenance.

Observation length

Building observation periods of any appreciable length using clinical data from commercial biospecimen providers posed a unique challenge. Patients were often seen once, or a few times over a short period of time for specimen collection, and provided self-reported medical information collected for a specific research purpose. Both aspects violate the longitudinal and observational intent of OMOP, but nonetheless could be modeled in the CDM using appropriate time windows in the observation_period table and careful type concept selections.

CONCLUSION

Data was structurally transformed to CDM specifications with minimal customizations, allowing downstream users to take advantage of the OHDSI suite of software, such as ATLAS to generate machine learning labels using cohort builder functionality, or tools to assess clinical data quality like the DataQualityDashboard. Semantic normalization of custom facts proved the most challenging aspect of this project, requiring significant cross-functional coordination between researchers and informaticists. Our approach demonstrates the OMOP CDM allows for novel use cases while maintaining its consistent definition.

REFERENCES

1. Zura R, Irwin DE, Mack CD, Aldridge ML, Mackowiak JI. Real-world evidence: A Primer. Journal of Orthopaedic Trauma. 2021 Mar;35(1). doi:10.1097/bot.0000000000002037
2. Ma C, Wang X, Wu J, Cheng X, Xia L, Xue F, et al. Real-world big-data studies in laboratory medicine: Current status, application, and future considerations. Clinical Biochemistry. 2020 Oct;84:21–30. doi:10.1016/j.clinbiochem.2020.06.014
3. Naidoo P, Bouharati C, Rambiritch V, Jose N, Karamchand S, Chilton R, et al. Real-world evidence and product development: Opportunities, challenges and risk mitigation. Wien Klin Wochenschr. 2021 Aug;133:840–6. doi:10.1007/s00508-021-01851-w. Epub
4. Hiramatsu K, Barrett A, Miyata Y. Current status, challenges, and future perspectives of real-world data and real-world evidence in Japan. Drugs - Real World Outcomes. 2021;8(4):459–80. doi:10.1007/s40801-021-00266-3
5. Papez V, Moinat M, Voss EA, Bazakou S, Van Winzum A, Peviani A, et al. Transforming and evaluating the UK Biobank to the OMOP common data model for COVID-19 research and beyond. Journal of the American Medical Informatics Association. 2022;30(1):103–11. doi:10.1093/jamia/ocac203

6. Michael CL, Sholle ET, Wulff RT, Roboz GJ, Campion TR. Mapping Local Biospecimen Records to the OMOP Common Data Model. AMIA Jt Summits Transl Sci Proc. 2020 May 30;422–9.