Mapping Local Biospecimen Records to the OMOP Common Data Model

Chelsea L. Michael, MS1,2, Evan T. Sholle, MS3, Regina T. Wulff, MS4, Gail J. Roboz, MD4, Thomas R. Campion, Jr., PhD2,3,5,6

1Department of Health Informatics, Memorial Sloan Kettering Cancer Center, New York, NY; 2Department of Healthcare Policy & Research, Weill Cornell Medicine, New York, NY; 3Information Technologies & Services Department, Weill Cornell Medicine, New York, NY; 4Department of Medicine, Weill Cornell Medicine New York, NY; 5Clinical and Translational Science Center, Weill Cornell Medicine, New York, NY; 6Department of Pediatrics, Weill Cornell Medicine, New York, NY

Abstract

Research to support precision medicine for leukemia patients requires integration of biospecimen and clinical data. The Observational Medical Outcomes Partnership common data model (OMOP CDM) and its Specimen table presents a potential solution. Although researchers have described progress and challenges in mapping electronic health record (EHR) data to populate the OMOP CDM, to our knowledge no studies have described populating the OMOP CDM with biospecimen data. Using biobank data from our institution, we mapped 26% of biospecimen records to the OMOP Specimen table. Records failed mapping due to local codes for time point that were incompatible with the OMOP reference terminology. We recommend expanding allowable codes to encompass research data, adding foreign keys to leverage additional OMOP tables with data from other sources or to store additional specimen details, and considering a new table to represent processed samples and inventory.

Introduction

Leukemia poses a substantial burden to the US population with 61,780 new diagnoses and 22,840 deaths projected for 2019.(1) About 1.5% of US men and women will have a leukemia diagnosis in their lifetime.(2) Within leukemia sub-types, acute myeloid leukemia (AML) will account for 21,450 new diagnoses and 10,920 deaths in 2019 projections.(1) Some AML patients will first develop myelodysplastic syndrome (MDS) before progressing to AML.(3) While precision medicine has dramatically improved outcomes for other leukemia sub-types, such as Gleevec for Philadelphia-positive chronic myeloid leukemia, the current AML standard of care includes only conventional treatments.(4) Targeted therapies for AML are under investigation,(4) but more research is needed to better understand the biology and molecular etiology of the disease.

To develop targeted therapies and to identify patients most likely to benefit from them, researchers must understand genetic variants, their associated pathways, and their role in oncogenesis. Such research depends in part on biospecimens collected from patients at various disease points. However, biospecimens alone are not sufficient to drive innovation in translational research. To best utilize samples, researchers need data about sample characteristics and inventory, but they also require increasingly rich, detailed clinical data about the patients whose samples are available.(5)

One potential approach to address the need for detailed clinical context for each biospecimen is to integrate biospecimen data with electronic health record (EHR) data. To promote EHR-data-driven research, a wide range of academic medical centers and pharmaceutical companies have adopted the Observational Medical Outcomes Partnership common data model (OMOP CDM). Maintained today by the Observational Health Data Sciences and Informatics (OHDSI) consortium, the OMOP CDM provides a common modality for mapping clinical data from both transactional/claims-derived data sets and EHRs to a universal, standardized format consisting of a relational data model with distinct tables for representing patients, visits, conditions, procedures, measurements, and specimens, among other domains.(6) The OMOP CDM also requires each site to map its terminologies (whether local or standardized) to standardized clinical reference terminologies – for example, sites that store diagnostic data using ICD-9/10 codes must map these diagnoses to the “standard” SNOMED-CT codes.(6) Several studies have reported successes and failures in mapping and loading drugs, conditions, procedures, and other clinical data from EHR and claims databases to the OMOP CDM.(7-12) However, to the best of our knowledge, no studies have described mapping biospecimen data to the OMOP CDM.
The goal of this study was to evaluate the feasibility of transforming data from a biospecimen database at our institution to the OMOP CDM to facilitate integration with EHR data to support AML and MDS research.

Methods

Setting

Weill Cornell Medicine (WCM) Physician Organization is a multi-specialty quaternary care center with 20 sites in New York City and 2 million patients. WCM’s 900 physicians have admitting privileges at its affiliated hospital, NewYork-Presbyterian Hospital (NYP). For each patient, WCM and NYP clinical systems share a unique medical record number (MRN). The WCM-NYP Leukemia Program is a field leader, treating a high volume of patients with a broad variety of hematologic malignancies and disorders, including AML and MDS. The Leukemia Program also conducts related basic, translational, and clinical research and coordinates studies with multiple participating institutions.

Leukemia Program Research Data Management

To support its robust research portfolio, the Leukemia Program maintains a central biobank with an associated Leukemia Program Database (LPDB).(13) When a biospecimen arrives to the biobank, research staff enter specimen-level data into the LPBD, including patient MRN, patient diagnosis, study protocol number, study time point, specimen source, specimen preservative, specimen collection time, specimen collection date, total specimen volume (in mL), and other data. Like many research databases, the LPDB uses local codes (e.g., “AML” and “MDS”) rather than reference terminologies.

The available data varies with specimen; for example, samples collected from external institutions do not have a corresponding WCM MRN. The most common Diagnosis field codes are “AML” and “MDS,” but other diseases (both malignant and non-malignant) are also available. Time Point has codes representing when a patient donated a sample, either in the context of the patient’s disease course (e.g., “Newly diagnosed” or “Remission”) or the study calendar for the research protocol under which they donated the specimen (e.g., “Day 100”). Study calendar time points do not represent the individual’s disease status, but rather where the patient falls within a specific research protocol. On “Day 100” of a research protocol, some patients may still have malignant disease, while others may be in remission. Time point data may also be unavailable or unreliable. The most common Specimen Source codes are “Peripheral blood” and “Bone marrow” and the most common Specimen Preservative codes for blood samples are “EDTA” and “Sodium heparin.” For controls, studies may collect samples from healthy individuals and/or collect healthy tissue samples from patients. The biobank staff also records data about samples after processing (e.g., number of aliquots and storage location) in the LPDB. Researchers seeking samples with specific characteristics, e.g. “all remission samples for female patients aged 60-85 with AML who have never received a bone marrow transplant, have an IDH-2 mutation, and have been treated with decitabine,” must identify these samples based on clinical data points captured in other research data systems, as well as the EHR, which complicates the process of identifying samples for secondary analysis.

OMOP CDM Implementation at WCM

WCM previously has implemented the OMOP CDM to support secondary use of EHR data.(14) OMOP includes the CDM Specification, Standardized Vocabularies and various tools to support OMOP implementers and users.(6) The CDM Specification includes table definitions, such as the Specimen table described in Table 1. OMOP uses Standardized Vocabularies for consistent data representation across implementations. Implementers use mappings to translate local codes or non-OMOP-preferred terminologies to the Standardized Vocabulary codes, called Concept IDs, permitted for each field. Concept IDs in each table serve as foreign keys to the OMOP Standardized Vocabularies tables. OMOP primarily uses SNOMED-CT as a Standard Vocabulary for concepts in the Specimen table. When a source code is unable to map to a standard, implementers use Concept ID = 0 to indicate no appropriate standard concept is available. Tables also include Source ID and Source Value fields to store raw (untransformed) data as they appear in the originating data source. When various systems or fields populate the same OMOP table fields, WCM has a local convention to use JSON annotations, including field labels for provenance, in the relevant Source Value fields to facilitate future use. The additional annotation necessitates increasing the character limits on
these optional fields. Finally, OMOP tools include Athena, which allows users to search and browse standard concept by OMOP field and table, and by reference terminology, hierarchies and codes.

**Table 1.** OMOP SPECIMEN table description, reproduced from the OMOP CDM v6.0 Specifications.(15)

<table>
<thead>
<tr>
<th>Field</th>
<th>Required</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>specimen_id</td>
<td>Yes</td>
<td>integer</td>
<td>A unique identifier for each specimen.</td>
</tr>
<tr>
<td>person_id</td>
<td>Yes</td>
<td>integer</td>
<td>A foreign key identifier to the Person for whom the Specimen is recorded.</td>
</tr>
<tr>
<td>specimen_concept_id</td>
<td>Yes</td>
<td>integer</td>
<td>A foreign key referring to a Standard Concept identifier in the Standardized Vocabularies for the Specimen.</td>
</tr>
<tr>
<td>specimen_type_concept_id</td>
<td>Yes</td>
<td>integer</td>
<td>A foreign key referring to the Concept identifier in the Standardized Vocabularies reflecting the system of record from which the Specimen was represented in the source data.</td>
</tr>
<tr>
<td>specimen_date</td>
<td>No</td>
<td>date</td>
<td>The date the specimen was obtained from the Person.</td>
</tr>
<tr>
<td>specimen_datetime</td>
<td>Yes</td>
<td>datetime</td>
<td>The date and time on the date when the Specimen was obtained from the person.</td>
</tr>
<tr>
<td>quantity</td>
<td>No</td>
<td>float</td>
<td>The amount of specimen collection from the person during the sampling procedure.</td>
</tr>
<tr>
<td>unit_concept_id</td>
<td>No</td>
<td>integer</td>
<td>A foreign key to a Standard Concept identifier for the Unit associated with the numeric quantity of the Specimen collection.</td>
</tr>
<tr>
<td>anatomic_site_concept_id</td>
<td>Yes</td>
<td>integer</td>
<td>A foreign key to a Standard Concept identifier for the anatomic location of specimen collection.</td>
</tr>
<tr>
<td>disease_status_concept_id</td>
<td>Yes</td>
<td>integer</td>
<td>A foreign key to a Standard Concept identifier for the Disease Status of specimen collection.</td>
</tr>
<tr>
<td>specimen_source_id</td>
<td>No</td>
<td>varchar(50)</td>
<td>The Specimen identifier as it appears in the source data.</td>
</tr>
<tr>
<td>specimen_source_value</td>
<td>No</td>
<td>varchar(50)</td>
<td>The Specimen value as it appears in the source data. This value is mapped to a Standard Concept in the Standardized Vocabularies and the original code is, stored here for reference.</td>
</tr>
<tr>
<td>unit_source_value</td>
<td>No</td>
<td>varchar(50)</td>
<td>The information about the Unit as detailed in the source.</td>
</tr>
<tr>
<td>anatomic_site_source_value</td>
<td>No</td>
<td>varchar(50)</td>
<td>The information about the anatomic site as detailed in the source.</td>
</tr>
<tr>
<td>disease_status_source_value</td>
<td>No</td>
<td>varchar(50)</td>
<td>The information about the disease status as detailed in the source.</td>
</tr>
</tbody>
</table>

**Data Collection**

We collected data from the LPDB on the most high-volume specimens, specifically those with an AML or MDS diagnosis sourced from a bone marrow or peripheral blood specimen that included a local WCM MRN. For peripheral blood specimen sources, we only selected records with EDTA or sodium heparin specimen collection preservatives. We selected the fields corresponding to these criteria, as well as fields for specimen ID, total volume and time point.

**Mapping to OMOP CDM**

We used Athena to determine standard concepts for the leukemia data. When one-to-one field mappings were inadequate, we considered many-to-one options, as well as drew on biobank knowledge (both general biobanking experience and knowledge of LPDB) to select the best data representation.
Evaluation

Based on the mapping rules we identified, we performed an extract, transform, and load (ETL) of the LPDB data into the OMOP Specimen table using structured query language (SQL) implemented in Microsoft SQL Server 2012. We analyzed how well the Specimen table represented the LPDB source data. Specifically, we measured the percentage of LPDB codes and records that mapped to the Specimen table; one-to-one compared to many-to-one mapping and one-to-many mappings; and where mapping lost detail or was not feasible.

Results

Mapping

When necessary, we combined multiple LPDB fields to map to the best OMOP Concept ID (Figure 1). For example, mapping only the LPDB Specimen Type field would have mapped to a more general blood specimen concept, such as “Peripheral blood specimen” (Concept ID = 4047495)(16) or “Blood specimen” (Concept ID = 4001225),(17) whereas a combination of LPDB Specimen Type and LPDB Specimen Collection Preservative allowed us to map to OMOP concepts of “Blood specimen with EDTA” (Concept ID = 40482922)(18) and “Blood specimen submitted in heparinized collection tube” (Concept ID = 40486989)(19) for the Specimen Concept ID field. In contrast, while we had specimen and preservative available for bone marrow, OMOP concepts were less granular and did not differentiate between preservatives. We simply mapped the LPDB Specimen Type to “Bone marrow specimen” (Concept ID = 4000623)(20) whether it was collected in tubes with EDTA, sodium heparin or another preservative. For other fields, we drew on domain-specific knowledge to identify the best concepts for mapping. For example, the LPDB data did not record an anatomic site (a required OMOP Specimen field) for blood draws, but “Peripheral blood” specimen types would be obtained through venipuncture. As OMOP had Procedure tables and SNOMED-CT codified a variety of relationship types, we used Athena to identify “Venous structure” (Concept ID = 4104340)(21) as the direct procedure site for “Venipuncture for blood test”, and chose that as the most appropriate OMOP Anatomic Site Concept ID. We used a similar approach to map bone marrow samples to “Bone marrow structure” (Concept ID = 4029619).(22)

OMOP Disease Status Concept ID was difficult to map based on the available LPDB data. The OMOP Disease Status Concept ID field designated the disease status of the individual sample, with three allowed concepts: “Normal” (Concept ID = 4069590),(23) “Abnormal” (Concept ID = 4135493)(24) and “Malignant” (Concept ID = 4066212).(25) The data set included only AML and MDS samples, which could have malignant or normal blood and bone marrow, depending on disease status at collection. We examined LPDB Time Points and mapped common values which indicated clear disease status that aligned with OMOP allowed disease status concepts. We did not map other disease status time points (e.g., “CRi,” which is the local code for complete remission with incomplete blood count recovery, and “Remission”) due to relatively low frequency in the database. We could not map samples with study calendar time points (e.g., “Screen,” “Day 100”) or blank time points. To facilitate later expansion to include additional codes, we used LPDB Specimen Source, Diagnosis and Time Point to map to Disease Status Concept ID (Figure 1), but within selected data set, only Time Point was essential.

We also needed to combine separate LPDB Date and Time fields for the required Specimen Datetime field. However, we mapped LPDB Date to Specimen Date and LPDB Volume to Quantity without manipulation. We generated Specimen ID using auto-number and we joined the LPDB data on MRN to the OMOP Person table to obtain the previously-generated Person ID for the patient who donated each sample.

Two required OMOP Specimen fields, Specimen Type Concept ID and Unit Concept ID, did not correspond to LPDB fields. Specimen Type Concept ID only had one allowed concept to indicate the data source was an “EHR Detail” (Concept ID = 581378), which was an OMOP-generated concept, rather than drawn from reference terminologies. The “EHR Detail” concept did not accurately represent our research data. Therefore, we mapped to Concept ID = 0. Unit Concept ID coded the unit for the sample quantity (i.e., sample volume for mapped LPDB records). Based on knowledge of the data set, we inferred the unit for the LPDB blood and bone marrow sample volume was mL.

The last five fields are optional Source ID and Source Value fields. We did not manipulate LPDB data to populate these fields. However, we followed the local WCM JSON convention (described above) when source data came from more than one LPDB field.
Figure 1. Mapping source fields in Leukemia Program Database (LPDB) to OMOP common data model fields for blood (red), bone marrow (yellow) and both (orange). Inferred data did not exist in LPDB source data; instead we used knowledge of the LPDB for those mappings. Lines indicate the source fields (or lack thereof) used to populate each OMOP Specimen field. Optional Source fields (omitted) have the same relationship as their corresponding standardized fields.
Population of OMOP CDM

Of the 5,453 AML and MDS records with local MRNs in the data set, we mapped 1,397 (26%) to the OMOP CDM. The remaining records (n = 4,056) did not map due to the required Disease Status field. Of 24 time point codes in the LPBD data set, we mapped only 5 (20.8%)—“Complete response,” “Relapse,” “Newly diagnosed,” “Baseline,” and “Residual disease”—to allowed Disease Status Concepts; the other 19 codes were study calendar time points, blank or codes with low frequency in the data set. For one field, Specimen Type Concept ID, we had to map to Concept ID = 0 to indicate no appropriate Concept ID was available.

We populated all 15 OMOP fields (required and optional) for each record. Of the 15 fields, we mapped LPDB data to populate 13 fields and used knowledge of the LPDB to populate two fields (Specimen Type Concept ID and Unit Concept ID). We hard coded mappings for all standard Concept ID fields (three from LPDB data, two from inferred data). Of the 15 fields, we used many-to-one mappings for three fields (Specimen Concept ID, Disease Status Concept ID, Specimen Datetime), and we did not manipulate the source data for two fields (Specimen Date and Quantity) in addition to the Source Value and Source ID fields, which are intended to store unmanipulated data. We used auto-number for the primary key field and used a join to the corresponding OMOP table in the ETL to populate one foreign key field.

Within the LPDB data, we used one field, Specimen Source, in mappings to three Concept ID fields (Specimen Concept ID, Anatomic Site Concept ID and Disease Status Concept ID). We also used Specimen Date for two fields (Specimen Date and Specimen Datetime). While OMOP captured specimen preservative for blood samples, it did not for bone marrow. We also lost time point detail in mapping to only three available OMOP disease statuses.

Discussion

Using a biospecimen database from one laboratory at an academic medical center, we successfully mapped 26% of source records to populate the OMOP Specimen table with an ETL. All remaining records did not map due to challenges related to the disease status. One field did not have any allowed concepts appropriate for the data. Overall, the mapping was not clean and straightforward; we combined source data fields (sometimes only for a sub-set of records) for some OMOP fields, as well as drew on the same source field to populate more than one OMOP field. The populated OMOP table was 100% complete, with data for all fields for every record.

While others have described OMOP implementations using EHR and claims data, to our knowledge this is the first study to describe population of the Specimen table with local biospecimen data. Our findings suggest areas where the Specimen table could be improved. Additional Concepts may be useful; for example, Specimen Type Concept ID only allows “EHR detail”, an OMOP-generated concept. However, other OMOP tables have “Research administration” (Concept ID = 44803476)(26) as an option, which would better fit our use case. Within the research and leukemia realms, additional disease statuses may be relevant (e.g., differentiating between samples at diagnosis, relapse and residual disease, rather than grouping them as “malignant”), though OHDSI must also balance specific group needs with general requirements.

Alternately, at the field level, the Specimen table could add a Visit Occurrence ID foreign key to the join to the Visit Occurrence table; research specimens are often collected with clinical samples which may have results available to supply additional nuanced data (e.g., residual disease vs relapse). Without that field, we may be able to join to the Visit Occurrence table using Person ID and Specimen Date or Specimen Date Time to leverage linked results to map additional records (i.e., those with calendar time points) or when additional details are needed for a project query. Ideally, we would also ETL data from other WCM biobank databases into the WCM OMOP Specimen table, which would introduce a need to identify the biobank laboratory housing each respective specimen. We could address this via data provenance (data from LPDB is about samples managed by the LPDB biobank team) and use JSON to annotate source values; alternately, each biobank could be conceptualized as a physically distinct Care Site in the corresponding OMOP table, and the Specimen table could include a Care Site ID foreign key.

Finally, the Specimen table captures specimens collected from the patient, not what is physically stored after processing (e.g., aliquots). Another table would be needed to adequately represent both clinical sample (slides and formalin-fixed paraffin-embedded tissue blocks (FFPE)) and research aliquots, as well as additional derivatives (slides cut from FFPE, DNA extractions) and their status (e.g., exhausted, sent for additional processing).
While common data models have many strengths including standardizing data to facilitate research at an institution or multi-institution level, our results demonstrate implementation challenges for research data, specifically biospecimens. Biospecimen data may be incomplete or local codes may not align well with the common data model vocabularies, as in the case of study calendar time points. Some fields could be easily manipulated (e.g. combining date and time fields and applying the appropriate format), while standard concepts required hard coding in the ETL (e.g., Specimen Concept ID, which combined two fields for blood samples).

We examined unmapped records to explore opportunities to expand the ETL. Within the selected data set, all remaining records failed mapping due to time point. Some codes used in low frequency like “Pre-transplant” (79/5,453 records, 1%), “CR” (76/5,453 records, 1%) and “Pre-induction” (12/5,453 records, 0.2%) could potentially be mapped. Majority of remaining time points were ambiguous and would require additional data sources (e.g., discrete pathology results data for clinical samples taken on same date, if available). For records we did not select, we do not expect to include the 932/8,523 (11%) records without an MRN. Of the 2,138 records with an MRN, 2,108 had other diagnoses, of which acute lymphoblastic leukemia (ALL) was the most common (528 records); however, only 16 of those ALL records had time points that could map to disease status. The 30 AML and MDS records we did not select had other sample types (e.g., 4 blood samples with other collection preservatives, 24 oral wash samples). We could map control samples (e.g. control sample types or normal diagnosis), which would not depend on time point; however, we would only gain 88 records. Overall, this emphasizes that the main challenge in mapping the LPBD data to the OMOP Specimen table was populating the disease status field. To expand our ETL to as many records as possible, we would need to explore additional data sources or to code the disease status as Concept ID = 0. If we were to use Concept ID = 0, each query to identify samples would need to incorporate additional fields to determine the disease status or returned records may require manual chart review.

In the current analysis, we focused on data from a single biobank. The Leukemia Program biobank was one of the more mature efforts at WCM, and its databases resulted from prior informatics collaborations. Therefore, certain LPDB characteristics may not be representative of other biobanks, but our experience may inform efforts at other institutions. Like many biobanks, available data depends on the study for which each sample was collected. Some samples come from outside institutions (and therefore will not have an MRN), and biobank staff depends on collaborating researchers to supply data. (Some collaborators may have lower motivation or resources available to enter high-quality data for future use, as their primary goal is their current study.) The hematologic malignancy focus also brings separate priorities and challenges which may not align with use cases that informed the OMOP Specimen table design. From our general biobanking experience, we suspect the OMOP Specimen table was designed from a solid tumor perspective, where tissue may be divided or macrodissected into tumor and normal samples and anatomic site is more relevant.

We implemented the OMOP Specimen table at WCM using leukemia biobank data as a novel way to improve biobank efficiency and utilization. We selected sample characteristics that were most common in the database to create our study data set. We created mappings for codes and fields, using varied approaches across codes and combinations of codes. Finally, we analyzed the resulting ETL, opportunities to both expand our ETL and potential considerations for future OMOP versions.

**Conclusion**

Our OMOP Sample table implementation demonstrates challenges within curating clinical research data in general, and for biospecimens in particular. We expect other biobank informatics teams to face similar challenges and trade-off decisions as they implement OMOP or other common data models. While imperfect and incomplete, the opportunity to combine limited data collected for biobanking with more comprehensive and standardized EHR datasets in a common data model dramatically increases the utility of samples for additional studies.

**References**


