What Gets Lost in Translation: Measuring Data Loss in OMOP Conversion

Authors

PJ Allen, Mike Elashoff, Andrew Howland, Clara Oromendia
Cornerstone AI
www.cornerstone.ai

Clara Oromendia
Chief Product Officer, Co-Founder at Cornerstone AI

Andrew Howland
Co-Founder at Cornerstone AI

Data that has been processed by Cornerstone AI:

Types

- EMR
- Specialty EMR (e.g. Oncology)
- Claims
- Labs
- EMR + Claims
- EMR + Labs
- EMR + Claims + Labs
- Registries
- Clinical Trials (SDTM/ADaM/Rave)
- OMOP
- App Data (Patient Reported Outcomes)
- Sensor Data

Diseases

- Alzheimer's Disease
- Breast Cancer
- Crohn's Disease
- COVID-19
- Diabetes
- Healthy Controls
- Kidney Disease
- Melanoma
- Multi-Indication Oncology
- Non-Small Cell Lung Cancer
- Orthopedics
- Prostate Cancer
- Rheumatoid Arthritis
- Sickle Cell Disease

Conclusion

The 90/10 rule for OMOP conversion: OMOP enables cross study research but some information is lost in translation.

Background

- Real-world data (RWD) come from diverse source schemas, such as unique EHR systems, registries, claims, etc.
- The OMOP Common Data Model standardizes RWD to support interoperability, analysis reproducibility, and cross-source data harmonization.
- OMOP's design prioritizes standardization and cross-source interoperability which can result in information loss or simplification as data is transformed into OMOP format.
- The goal of this work is to quantify and characterize information loss during OMOP conversion.

Data

- 502 breast cancer patients from an EHR derived oncology dataset.
- Includes biomarker, biopsy, comorbidities, demographics, diagnoses, disease history, medication, labs, line of therapy, therapy response, metastases, procedure and vitals data.

Methods

- Data is processed using the Cornerstone AI (CSAI) platform, trained on billions of clinical data records spanning a variety of disease areas and types.
- CSAI's OMOP Conversion module automates mapping, standardization, and schema conversion, and lets users further refine AI-generated results through an intuitive UI.
- After conversion, data loss was assessed as the proportion of "analytically useful" source data that could be mapped to OMOP at:
  - overall
  - patient-level
  - field-level
  - code-level
- Examples of fields deemed to be not analytically useful include: tracking ids, empty fields, or fields duplicating information found in other fields.

How It Works: Schema Mapping

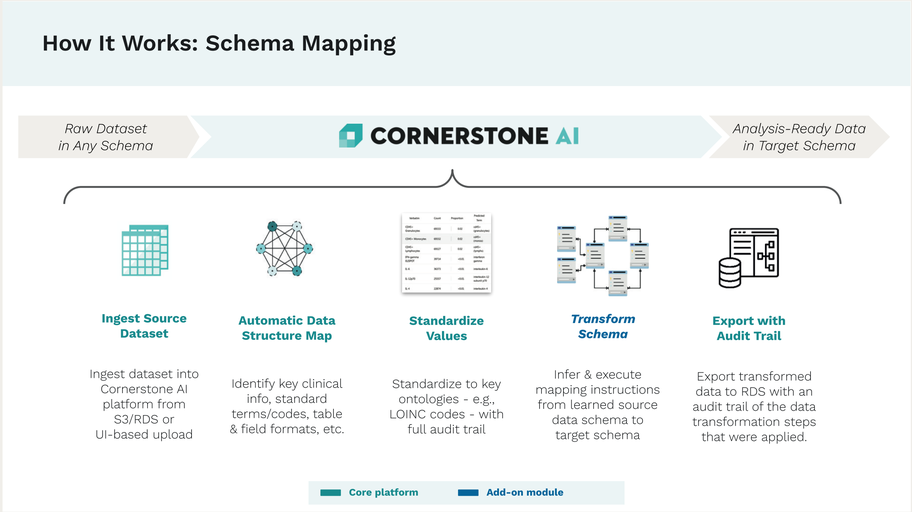

Figure 1. Overview of Cornerstone AI's OMOP Conversion Methodology

Results

Field-Level & Overall Mapping Rates

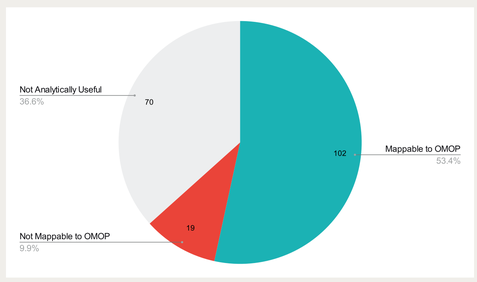

Figure 2. Field-Level Mapping to OMOP

- Of 191 fields in the source dataset, 121 (63.3%) were deemed analytically useful (Figure 2).
- Of these 121 useful fields, 102 (84.3%) could be mapped to OMOP (Figure 2).
- When weighted by the number of observations per useful field, **91.6% of data points could be successfully mapped to OMOP.**

OMOP Transformation Example

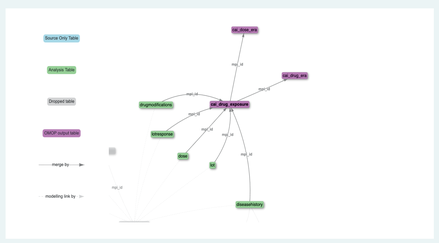

Figure 3. OMOP Drug Exposure: Source Table Mapping
Screenshot from Cornerstone AI Platform

OMOP Table	OMOP Field Name	Source Field Input ({table}_{field})
cai_drug_exposure	Subject	["dose_Subject","drugmodifications_Subject","lot_Subject", "lotresponse_Subject","diseasehistory_Subject"]
cai_drug_exposure	refills	"dose_refills"
cai_drug_exposure	stop_reason	"drugmodifications_change_type"
cai_drug_exposure	drug_exposure_end_datetime	["dose_drug_exposure_end_date","drugmodifications_change_date", "lot_end_date","lotresponse_lot_end_date"]
cai_drug_exposure	quantity	"dose_dose_value"
cai_drug_exposure	verbatim_end_date	["dose_drug_exposure_end_date","drugmodifications_change_date", "lot_end_date","lotresponse_lot_end_date"]
cai_drug_exposure	drug_exposure_end_date	["dose_drug_exposure_end_date","drugmodifications_change_date", "lot_end_date","lotresponse_lot_end_date"]
cai_drug_exposure	dose_unit_source_value	"dose_dose_units"
cai_drug_exposure	drug_exposure_start_datetime	["dose_drug_exposure_start_date","lot_start_date"]
cai_drug_exposure	drug_source_value	["dose_generic_name","dose_brand_name","drugmodifications_generic", "lot_regimen","lot_lot","lotresponse_regimen"]
cai_drug_exposure	drug_exposure_start_date	["dose_drug_exposure_start_date","lot_start_date"]
cai_drug_exposure	drug_source_concept_id	["dose_generic_name","dose_brand_name","drugmodifications_generic", "lot_regimen","lot_lot","lotresponse_regimen"]
cai_drug_exposure	drug_concept_id	["dose_generic_name","dose_brand_name","drugmodifications_generic", "lot_regimen","lot_lot","lotresponse_regimen"]
cai_drug_exposure	visit_occurrence_id	["dose_visit_occurrence_id","drugmodifications_visit_occurrence_id" ,"lot_visit_occurrence_id","lotresponse_visit_occurrence_id", "diseasehistory_visit_occurrence_id"]

Figure 4. OMOP Drug Exposure Table: Source Field Mapping
Note: The following OMOP Drug Exposure fields did not have data in the source: 'sig','lot_number','route_source_value','provider_id','drug_exposure_id','days_supply', 'drug_type_concept_id','visit_detail_id','route_concept_id'.

Code-Level Mapping Rates

- 9 OMOP coded fields had all three of 'source_value_name', 'source_concept_id', and 'concept_id'.
- CSAI auto-standardization improved mean mapping rates from 62 to 79% (71 to 85% frequency-weighted).
- Largest gains were found in the observation and measurement fields.

Table 1. Top 3 Fields with Greatest Concept Mapping Improvement Based on Unique Values

			Code Mapping Rate	
Table	Field	n	Before CSAI	After CSAI
observation	observation_source_value	7	0%	57%
measurement	unit_source_value	82	29%	80%
measurement	source_value	85	39%	66%

Table 2. Top 3 Fields with Greatest Concept Mapping Improvement Based on Frequency-Weighted Values

			Code Mapping Rate	
Table	Field	n	Before CSAI	After CSAI
observation	observation_source_value	3,090	0%	65%
measurement	measurement.source_value	164,906	61%	99%
measurement	measurement.unit_source_value	162,084	57%	82%

Selected References

[1] OHDSI. OMOP Common Data Model Documentation. U.S. FDA. (2021). Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products: Guidance for Industry. https://www.fda.gov/media/152503/download
[2] Voss EA et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc. 2015;22(3):553–564. https://doi.org/10.1093/jamia/ocu023

CORNERSTONE AI