

## Supplementary Material

### Supplementary Appendix A: Iterative QA Error Analysis Process Examples

#### Transformation Layer

Row count (completeness QA element):

- Patient Chemistry Laboratory Tests:
  - First QA run: missing 635,376 rows.
  - Error: *the QA script had an error and did not properly mimic ETL transform behavior. Specifically, the QA script was looking for records prior to October 1, 1999 and the ETL script correctly excluded them.*
  - Fix: *add logic/exclusion criterion to QA script and rerun row count check.*
  - Second QA run: *no errors.*
- Patient Outpatient Pharmacy Fill Records:
  - First QA run: 1,427 rows too many.
  - Error: *pharmacy fill records that were erroneous entries or deleted records that should have been filtered out were loaded into the transformation layer.*
  - Fix: *ETL patch to filter for erroneous data entry in fill records.*
  - Second QA run: *no errors.*
- Outpatient Administrative Procedure Codes:
  - First QA run: 2,357 rows too many.
  - Error: *rows that were erroneous or deleted in source data were retained in the transform.*
  - Fix: *ETL patch to filter out erroneous records.*
  - Second QA run: *no errors.*

#### Load Layer

Row count (completeness QA element):

- Patient Chemistry Laboratory to OMOP Measurement:
  - First QA run: 1,025,399 rows too many.
  - Error: *this was an incremental ETL error. As part of quarterly OMOP meta-data updates, the target (standardized) concepts for these rows changed domains, and the incremental logic did not catch the change and move the row from one OMOP table to the other (retained in measurement table).*
  - Fix: *incorporation of tracking all rows that are impacted by an OMOP meta-data row change in the CONCEPT, CONCEPT\_ANCESTOR, or CONCEPT\_RELATIONSHIP tables, to recompute impacted rows and update OMOP load.*
  - Second QA run: *no errors.*
- Outpatient Administrative Procedure Records to OMOP Procedure Occurrence
  - First QA run: 3,422 rows too many.
  - Error: *source concepts for these rows used to map to two concepts. OMOP meta-data updates revised the source-*

*to-target mapping to a single concept, but this incremental change was not captured.*

- Fix: *incorporation of tracking all rows that are impacted by an OMOP meta-data row change in the CONCEPT, CONCEPT\_ANCESTOR, or CONCEPT\_RELATIONSHIP tables, to recompute impacted rows and update OMOP load.*
- Second QA run: 2,939 rows too many.
- Error: *intermediate mapping table had not been fully updated to incremental change capacity as noted above.*
- Fix: *mapping table ETL patch to incorporate OMOP meta-data table changes as well.*
- 3<sup>rd</sup> QA run: *no errors.*

Referential Integrity (Relational conformance QA element):

- Patient Chemistry Laboratory tests to OMOP MEASUREMENT
  - First QA run: 9,983 rows with errors in PERSON\_ID in OMOP MEASUREMENT table.
  - Error: *old PERSON\_IDs—no longer in person table (orphan PERSON\_ID records).*
  - Fix: *add ETL code that checks for changed PERSON\_ID. This section of code had been omitted. Transform layer Patient Chemistry Laboratory Test table had been one of the last datasets that had been added and full incremental logic handling had not been added.*
  - Second QA run: *no errors.*
- Patient Chemistry Laboratory Tests to OMOP MEASUREMENT
  - First QA run: 28,910,497 rows with errors in PROVIDER\_ID in OMOP MEASUREMENT table.
  - Error: *old PROVIDER\_IDs—no longer in provider table (orphan PROVIDER\_ID records).*
  - Fix: *add ETL code that checks for changed PROVIDER\_ID. This section of code had been omitted. Transform layer Patient Chemistry Laboratory Test table had been one of the last datasets that had been added and full incremental logic handling had not been added.*
  - Second QA run: 3,299 rows from 8 PROVIDER\_IDs.
  - Error: *we were unable to determine why the patch failed to work. We manually forced recomputation of the impacted rows, and then resumed the incremental process.*
  - 3<sup>rd</sup> QA run: *no errors.*

Partial/multiple mapping (Value conformance QA element):

- OMOP MEASUREMENT table
  - First QA run: *50 source values (ICD10 codes) have more than one SOURCE\_CONCEPT\_ID.*
  - Error: *we changed our logic from ICD10 to ICD10CM to conform to United States standards. This change was not*

made for the source outpatient diagnosis table. All of these codes came from that single transform layer table.

- Fix: added ICD10CM mapping instead of ICD10 mapping to the Outpatient Administrative Diagnosis Code table ETL Load logic.
  - Second QA run: no errors.
2. OMOP DRUG\_EXPOSURE table
- First QA run: 52 source values partially mapped to SOURCE\_CONCEPT\_IDs.
  - Error: ETL missed these updates. They were previously unmapped NDC/VA Product Code medication mappings that were mapped in the OMOP meta-data update, and the incremental logic did not detect the changes.
  - Second QA run: 47 (mapping table was not updated for these rows).
  - Error: some of the errors in the original error batch had not been flagged to be updated in the patch. We forced manual recomputation of those rows.
  - 3<sup>rd</sup> QA run: no errors.

## Supplementary Appendix B: Top 10 questions to consider when incrementally transforming data into a CDM

QA question	QA element	QA resolution
1. Were data inclusion and exclusion criteria applied consistently to source data?	Completeness	Row counts between source and transformation levels
2. Have all new and deleted source rows been accounted for in the transformation?	Completeness	Row counts between source and transformation levels
3. Were CDM source-to-target mapping updates uniformly applied to transformed data?	Completeness	Row counts between transformation and OMOP levels
4. Do primary key/foreign key relationships in the OMOP tables align in the same way they do in source data?	Relational conformance	Referential integrity checks between transformation and OMOP levels
5. Do any foreign keys exist in fact tables that do exist as primary keys in a parent table?	Relational conformance	Orphan checks between OMOP tables
6. If a source concept has a standard mapping, are all instances of the source concept mapped?	Value conformance	Partial/multiple mapping of each OMOP table
7. Do any nonstandard concepts exist as Concept IDs in OMOP fact tables?	Value conformance	Concept constraints checks of each OMOP table
8. Do any source concept IDs have more than one relationship with target?	Value conformance	Partial/multiple mapping of each OMOP table
9. Do all source concepts have one and only one source concept ID?	Value conformance	Partial/multiple mapping of each OMOP table
10. Does each domain contain only concepts that belong to that domain?	Value conformance	Domain ID checks of each OMOP table

Abbreviations: CDM, common data model; OMOP, Observational Medical Outcomes Partnership; QA, quality assurance.