N3C Data Quality Gates Focus Group, #3

11May2020

Phenotype and Data Acquisition
Friday Weekly Meeting
15May20

Data Quality: Site-Side

File Spec Conformance

- A straightfoward “quality” aspect: conformance to our file spec.
- Python script in progress to help sites consistently export tables in our chosen format. Optional to use, but hopefully helpful.
- Our thought is to create a file confirmance checking script to run prior to submission. Would require Python.
- As we have not written this yet, it might be worth seeing what the variability is among early sites. (If it ain’t broke…)

A Quandary

- Data quality is much more complicated!
- Harmonization team has been surveying different data QA techniques across data models.
- Some DQ techniques require significant time/compute resources. (The PCORnet data curation, for one.)
- Some DQ checks are much more strict than others.
- Sites are presumably already running required data checks in their CDMs (with or without N3C). Do we ask them to run the checks again?

Options

- Option 1:
  - All sites run their data model-specific DQ checks locally prior to every submission to N3C, remediate problems, and send “clean” DQ check report to N3C along with data payload.
  - “Perfect world” option—likely not sustainable.
  - Too much variation between different models’ requirements.
  - Maybe too much information for Harmonization to consume.
  - Verdict? ☹

Options

- Option 2:
  - All sites run their data model-specific DQ checks locally on some recurring schedule (monthly? quarterly?), and send “clean” DQ check report to N3C when run.
  - Still have the problem of variation between different models’ requirements.
  - Less information for Harmonization to consume—still quite a bit.
  - Verdict? 😐

Options

- Option 3:
  - N3C defines a ***minimum*** set of our most-important site-side checks—same checks for each model, just different syntax.
  - Phenotype and Data Acquisition group writes SQL for each model to execute these checks natively in the database (i.e., not with a third-party tool).
  - Quality checks should run in a reasonable amount of time—will require testing.
  - Sites would be expected to correct errors prior to sending—must submit “clean” DQ check table with each data payload.
  - Can produce structured, machine-readable information for Harmonization to consume.
  - Verdict? 😊 (?)

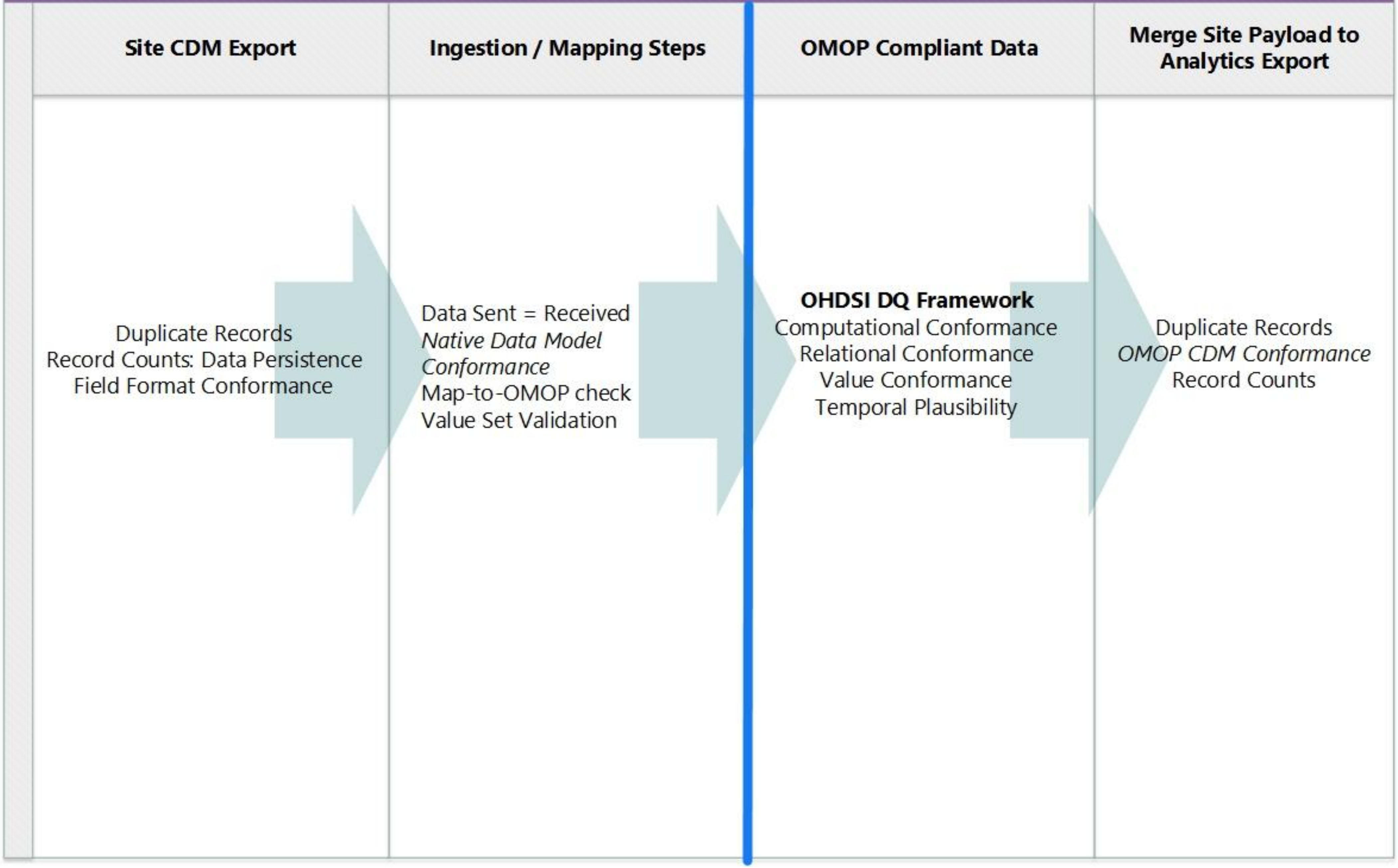
Timeline

- This will take coordination between Data Acquisition team and Harmonization team.
- Question: Should we hold up data submission until these checks are written? Or implement this gradually?

Proposal

- Create Minimum DQ Gate configuration: v0.1
  - Architecture proof of concept
- Include Phenotype & DA export script in DQ Gate config
- Eliminate Duplicate / Overlap DQ testing
- Minimum necessary tests, min threshold @ each step

N3C Data Quality Gates Minimum Configuration: Minimum DQ Duplication



Site CDM Export	Ingestion / Mapping Steps	OMOP Compliant Data	Merge Site Payload to Analytics Export
Duplicate Records Record Counts: Data Persistence Field Format Conformance	Data Sent = Received *Native Data Model Conformance* Map-to-OMOP check Value Set Validation	**OHDSI DQ Framework** Computational Conformance Relational Conformance Value Conformance Temporal Plausibility	Duplicate Records *OMOP CDM Conformance* Record Counts

Questions from last meeting *(resolution?)*

- Where are there redundant DQ tests?
  - Where should these persist / are necessary? Can any be "pruned?"
  - *Minimum Redundancy & coverage*
- What are the (min / max) DQ tests that should be performed in the ingestion / mapping phases? Includes: accommodation for differences in CDMs
  - *Minimum but necessary DQ checks*
  - *Eliminates accommodation for CDMs*
- What are the DQ tests that should be leveraged in the OHDSI DQ toolkit?
  - What thresholds should be set / managed?
  - What are the expected outcome(s) associated with setting thresholds?
  - *Minimum necessary established in remaining DQFG meetings*
- What DQ testing should be created for the merge step?
  - *Suggesting: Duplicate Records, Counts, OMOP conformance (?)*
- De-duplication of patients - strategies
  - Utilization of Hashes
  - *Accommodated for / created in Phenotype & DA export script*

Remaining Discussions

- Phenotype & Data Acquisition Gate
  - Duplicates
  - Counts
  - Field conformance
- Data Ingestion & Harmonization Gate
  - Native CDM conformance – necessary?
  - Map Validation: Adeptia functionality
  - Value Set validation: Adeptia / API to Athena
- OHSDI Gate
  - Minimum value / conformance config
  - Threshold default / minimum?
- Data Merge Gate
  - Counts: incrementing as expected
  - Duplicate patients – is this possible?
  - OMOP conformance: is this necessary?