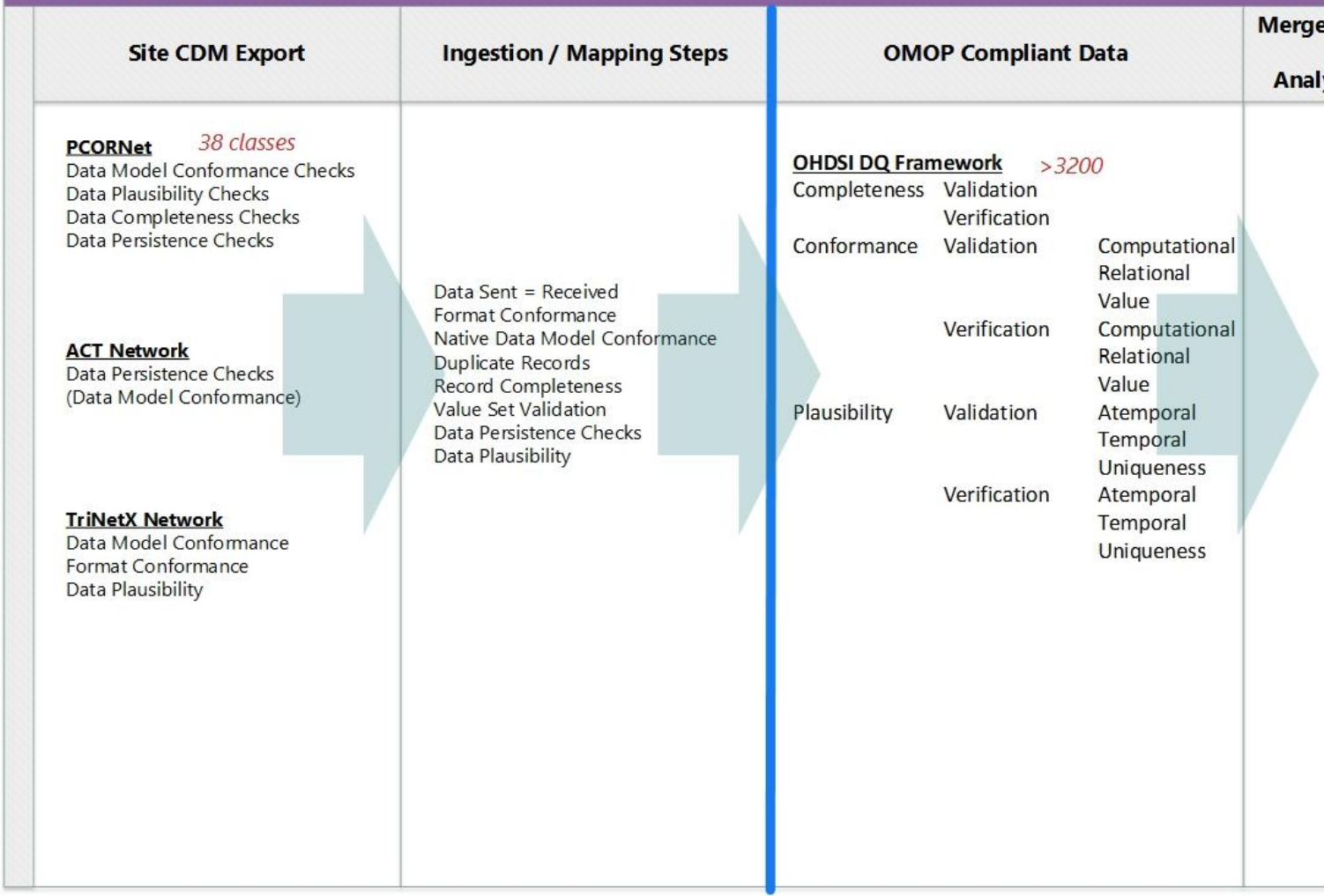
N3C Data Quality Gates Focus Group #2

8May2020

Emerging DQ Principles

1) Utilizing R for DQ scripting for ingestions & mapping steps
2) Tighter data controls / narrow variance on COVID19 Phenotype definitions
    a) Correct LOINC codes
    b) Attention to comment / text fields
3) Other data “corrections’ not in scope
    a) Persist source data in transformations and pass along to analytics
4) Preserve semantic quality through curated Value Set mapping
5) Not all sources will have the complete sets of data classes for every record
6) Detailed mapping review exposing quality & phenotype details

N3C Data Quality Gates



Questions

1) Where are there redundant DQ tests?
    a) Where should these persist / are necessary? Can any be “pruned?”
2) What are the (min / max) DQ tests that should be performed in the ingestion / mapping phases? Includes: accomodation for differences in CDMs
3) What are the DQ tests that should be leveraged in the OHDSI DQ toolkit?
    a) What thresholds should be set / managed?
    b) What are the expected outcome(s) associated with setting thresholds?
4) What DQ testing should be created for the merge step?
5) De-duplication of patients - strategies
    a) Utilization of Hashes