

An Automated Data Quality and Curation Pipeline for Producing GenoSphere OMOP Common Data Model Clinico-Genomic Datasets for the Helix Research Network

**Lance Eighme, Lisa M McEwen, Xavier Araujo, Kwasi Agyeman, Anna Swigart
Helix, Inc., 400 S El Camino Real Suite #300, San Mateo, CA 94402, United States**

Background

The Helix Research Network (HRN) is a collaborative alliance of healthcare organizations partnering with Helix to accelerate precision medicine through large-scale genomic profiling. By pairing population-scale genomic sequencing with longitudinal clinical data, the HRN builds North America's largest clinico-genomic ecosystem to drive biomedical discovery that we call GenoSphere™. We are currently at over 500,000 clinico-genomic records and on pace for 1M over the next 2 years. The quality of observational research depends fundamentally on the quality of the underlying data, and as part of the Helix Research Network (HRN), several health systems are adopting the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) to standardize their electronic health record (EHR) data. These data are used both internally and externally to support novel clinico-genomic studies. Similar to others in the OHDSI community, we want to ensure our data is of high quality to support accurate research and analytics.

As health systems increasingly adopt the OMOP Common Data Model to support large-scale analytics, the challenge of ensuring data quality across heterogeneous electronic health record sources has become critical.^{1,2} Despite the availability of community tools such as ACHILLES and the Data Quality Dashboard (DQD), significant gaps remain in operationalizing continuous, automated, and proactive data curation workflows that address source-level variability in coding practices, units of measure, and concept mapping.³ These challenges are compounded in multi-site research networks, such as HRN, where harmonization across diverse health system partners is essential for generating reliable evidence.

We present an automated data quality and curation pipeline that produces a highly validated GenoSphere "gold" level OMOP dataset through systematic identification, normalization, and resolution of data quality issues. This work addresses the OHDSI community's ongoing need for reproducible, scalable approaches to data quality management that go beyond detection to active remediation. Our pipeline leverages cloud infrastructure (AWS Glue) and incorporates artificial intelligence to support curation rule generation, representing a practical integration of automation, partner collaboration, and AI-assisted data stewardship within the OMOP ecosystem.

Methods

We developed an end-to-end automated data quality and curation pipeline built on AWS Glue⁴, designed to process EHR data from multiple health system partners into OMOP CDM format with enhanced quality assurance.

Core Data Element Definition: We defined a set of “core” data elements targeted for the highest level of quality validation and curation. These elements were selected based on their importance for downstream analytic use cases including clinical characterization, population-level effect estimation, and patient-level prediction.

Curation Steps: The pipeline implements the following curation processes:

- **Source code standardization:** Mapping source codes to OMOP standard vocabulary concepts with AI-assisted disambiguation.
- **Concept normalization:** Re-mapping already-mapped OMOP standard concepts to a preferred target vocabulary using OMOP concept relationships, reducing variability in how equivalent data elements are represented across network partners and improving cross-site consistency.
- **Units of measure normalization and conversion:** Converting disparate units to canonical representations to ensure cross-site comparability.
- **Unit of measure inference:** Applying rule-based and AI-driven logic to infer missing or ambiguous units from contextual data patterns.
- **Normalized data element frequency evaluation:** Assessing the distribution and completeness of data elements across network partners to identify systematic gaps or anomalies.
- **Curation filtering:** Records are filtered to retain only measurements with preferred units and expected result values. Measurements with mismatched units, missing results, or extreme outlier values indicative of data entry errors are excluded to reduce noise and improve data usability for downstream analytics.

Collaborative Feedback Loop: We maintain close collaboration with health system partners to iterate on data quality findings, supporting discovery and extraction of additional data elements from the EHR alongside standardization improvements.

Curation Application: The workflow is supported by a purpose-built application featuring an automated feedback loop where curation rules can be easily created or altered. The application integrates AI support for generating new inference and normalization rules and for mapping source codes to standard vocabulary terms.

Automated Reporting: All findings are automatically summarized through an automated reporting process, providing transparency and facilitating partner engagement.

Results

Our automated data quality and curation pipeline processed 1.25 billion input measurement records across multiple health system partners. The pipeline's impact is summarized across two phases: transformation and exclusion (Figure 1, Figure 2).

Transformation Phase: Source code standardization modified 178 million records (14.3%), harmonizing heterogeneous coding representations across sites. Measurement concept normalization had the largest transformation impact, re-mapping 284 million records (22.7%) to preferred vocabulary concepts through a combination of OMOP concept relationships and well-defined normalization rules applied per data element concept group, substantially reducing representational variability. Unit inference successfully resolved missing or ambiguous unit_concept_ids for 43 million records (3.5%), improving measurement completeness without manual intervention.

Exclusion Phase: Three independent exclusion criteria were applied to filter records that would otherwise introduce noise into downstream research and analytics. Missing values accounted for the largest exclusion at 41 million records (3.3%), followed by unit mismatches at 26 million records (2.1%), and extreme outliers at 2.7 million records (0.2%). The combined union of all exclusion criteria removed 59.2 million records (4.7%), indicating approximately 11.8 million records were flagged by more than one criterion.

The OMOP harmonized GenoSphere dataset retains 95.3% of targeted records, representing a high-fidelity, curated dataset suitable for clinical characterization, population-level effect estimation, and patient-level prediction. The substantial impact of concept normalization underscores the degree of variability present in source OMOP mappings across network partners, validating the need for this additional harmonization layer.

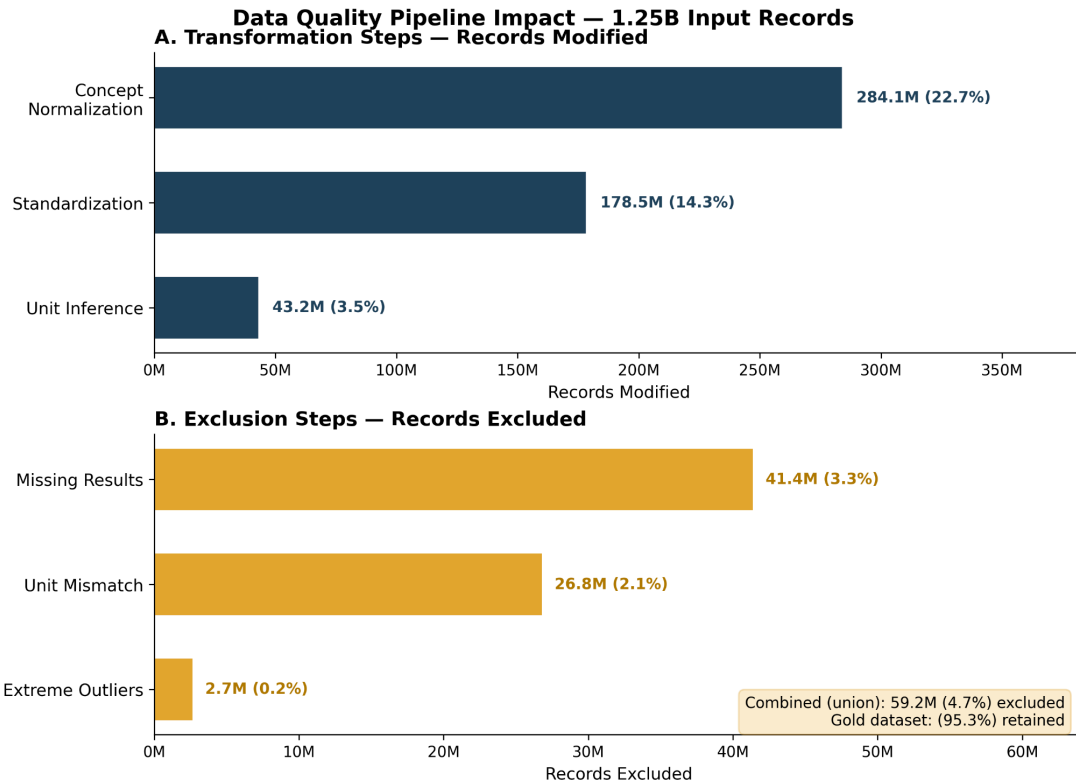


Figure 1: Data quality pipeline impact: records modified and excluded by curation step across 1.25 billion input records.

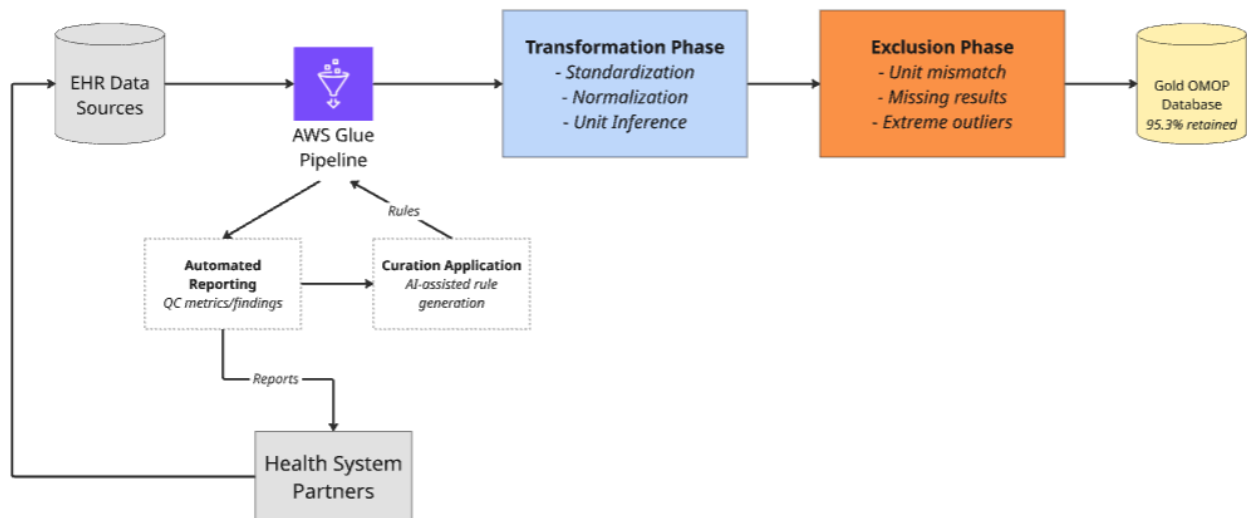


Figure 2: From EHR data to gold-level OMOP dataset through automated curation and partner collaboration

Conclusion

We demonstrate a scalable, automated data quality and curation pipeline that transforms raw EHR data into the GenoSphere OMOP CDM dataset through systematic standardization, normalization, and AI-assisted curation. Processing over 1.25 billion records across multiple health system partners, the

pipeline produces a 95.3% retention gold dataset while addressing substantial representational variability through automated transformation and filtering. The tight integration of automated quality checks, partner collaboration, and an intuitive curation application with AI support represents a reproducible framework that the OHDSI community can adopt to elevate data quality beyond passive detection toward active, continuous remediation. Future work will focus on expanding AI-driven rule generation and extending the pipeline to additional data domains.

References

1. Kahn MG, Callahan TJ, Barnard J, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS (Wash DC)*. 2016;4(1):1244.
2. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform*. 2015;216:574-8.
3. Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. *J Am Med Inform Assoc*. 2021;28(10):2251-2257.
4. AWS Glue. Amazon Web Services (AWS). Available from: <https://docs.aws.amazon.com/glue/>