

The Role of Data Operations in Observational Health Data Quality – A Landscaping Review

Katy Sadowski^{1,2}, Paul Nagy, PhD¹

¹Biomedical Informatics and Data Science, Johns Hopkins University School of Medicine, Baltimore, MD, USA, ²Boehringer Ingelheim Pharmaceuticals, Ridgefield, CT, USA

Background

Observational research is becoming increasingly prominent as a means of generating evidence about patient outcomes and healthcare quality. Utilizing real-world data (RWD) to generate real-world evidence (RWE) is far less resource-intensive than traditional research methods like randomized clinical trials.

However, RWD like electronic health records and administrative claims are not collected with research use in mind and thus have inherent limitations to their reliability and relevance for RWE generation.¹ Furthermore, extensive processing is often required to make RWD usable for research. Despite their importance in facilitating RWE research, data operations tasks (including extraction from source systems and transformation to analytic data models) introduce new opportunities for data loss and corruption. We can thus distinguish two major categories of potential data quality issues in retrospective observational research: those *inherent* to a data source and present from the time of data generation, and those *acquired* through data operations. Acquired quality issues can result from technical errors, such as bugs in code, or from design decisions impacting how data is modeled for analysis.

Acquired data quality deserves attention for several reasons. First, data processing errors and modeling assumptions can introduce systematic bias in RWE studies.²⁻⁷ Such bias can pass unnoticed when researchers lack access to source systems and when processing pipelines lack transparency. Second, acquired quality issues are preventable. While managing inherent data quality focuses on monitoring and reporting, managing acquired data quality involves preventing and correcting technical errors and validating design decisions. These distinct tasks have specific technology and workflow requirements which should not be conflated. Third, common data models (CDMs) like the Observational Medical Outcomes Partnership (OMOP) CDM are growing in prominence as a standard for RWE research. When implemented correctly, CDMs can significantly improve the quality of RWE.⁸⁻¹¹ However, transforming source data into a CDM is a complex, error-prone task. To realize the promise of CDMs, it is necessary to properly control for the quality risks introduced by the standardization process.¹²

This study examines the extent to which published RWE frameworks and guidance documents address the quality risks of data operations and assesses current practice through literature review and expert interviews. We propose solutions where gaps exist.

Methods

The scope of this review was limited to retrospective observational research involving structured electronic health record (EHR) and claims data in the United States and Europe.

A scoping literature review was performed which aimed to identify (1) frameworks and guidance documents addressing RWD curation, transformation, and quality assessment, and (2) examples of these activities in practice. First, Google and agency websites were used to search for relevant publications from 7 regulatory agencies; 8 non-governmental organizations; and 6 observational research networks.* Then, ChatGPT Deep Research was used to identify additional academic and grey literature.¹³ Deep Research was chosen for this task with the knowledge that it would not meet the requirements of a traditional systematic review; its strengths in surfacing prominent public-domain resources and grey

literature better suited our study's goals. Identified publications were screened for relevance; categorized; and reviewed to identify if and how they address the impact of data operations on data quality.

Expert interviews were conducted to assess how the quality risks of data operations are mitigated in current practice. Interview candidates were selected from organizations using the OMOP, PCORNet, or PEDSNet CDMs for observational health research. The selected candidates were responsible for data curation and/or transformation tasks at their respective institutions, which spanned academia and industry. Each participant attended a 30-minute interview in which they were asked to describe the data lifecycle for a research dataset under their management and the methods and tools used to monitor data quality at each lifecycle stage. Interview responses were analyzed qualitatively.

Results

The literature review identified 45 RWE frameworks and guidance documents and 45 published implementation examples.** We conducted eight expert interviews.

Frameworks & Guidance Documents

The reviewed RWE frameworks and guidance documents cover a variety of use cases and perspectives, from pivotal regulatory decision-making to collaborative public health research. Almost all explicitly address data operations quality. The following common themes emerged from our review:

- **Governance and process maturity** – well-defined operating procedures and adherence to standards for data integrity and interoperability are foundational requirements for producing a high-quality observational research dataset
 - *Example: Application of ISO standards to ensure data and software systems follow industry best-practices for process and quality management*
- **Documentation** – each step of the data lifecycle must be clearly documented to enable reproducibility and to provide a basis for quality assurance
 - *Example: Maintaining a human-readable description of all steps taken to transform source data into a common data model for research*
- **Transparency** – documentation must be made available to researchers to ensure design decisions and limitations of a dataset are known
 - *Example: Publishing data transformation documentation so that it may be reviewed by data users*
- **Quality control** – data quality must be monitored and provenance must be maintained at each step of the data lifecycle
 - *Example: Running completeness, conformance, and plausibility data quality checks on both the source data and the final research dataset*

The reviewed frameworks generally stop short of prescribing specific tools or techniques for operationalizing the above principles. Challenges introduced by cross-institution collaboration and data sharing are also frequently cited, but specific guidance for aligning quality standards across handoffs is lacking.

Implementation Examples

The reviewed implementation examples include research network documentation, academic

publications, and vendor white papers. These publications commonly describe using the quality of a research dataset as a proxy for the quality of its upstream curation and transformation processes – but they also note the difficulty of tracing quality issues to their root cause when doing so. Some publications describe easing this burden by performing source data characterization; implementing audit logging and inline quality checks; and tracking record-level provenance. The landscape remains fragmented, however, with most resources addressing specific data sources or transformation stages rather than a unified approach for end-to-end quality monitoring.

Interviews

Most interviewees worked with EHR data at academic medical centers. The most resounding theme across interviews was that the bulk of quality control occurred at the final stages of the data lifecycle. Source data capture, storage, and extraction were largely outside interviewees' control; proprietary or poorly documented source systems were frequently noted as posing quality risks. All interviewees ran their CDM's standard quality checks after transformation; some had also implemented unit tests, source-to-target comparisons, or traceability mechanisms. A lack of standard tools for upstream quality control was cited as a barrier to broader adoption. Several interviewees noted that it was common for researchers to identify extraction- or transformation-related quality issues at analysis time.

Conclusion

While real-world data quality guidance in the literature is comprehensive and widely acknowledges the importance of end-to-end quality monitoring, the high-level principles outlined in most RWE frameworks have yet to be translated into a unified implementation standard. Published examples and practitioner experiences reveal that quality control remains concentrated at the post-transformation stage, with limited tooling, proprietary source systems, and the absence of shared standards constraining upstream monitoring.

We propose three complementary areas of innovation to address this gap:

- **Open standards** – machine-readable data dictionary and quality check specifications applicable regardless of data model, pipeline stage, or transformation architecture
 - *Example: Extension of data quality tools like OHDSI's DataQualityDashboard to be agnostic to the input data model¹⁴*
- **Technology** – AI-enabled tooling to translate quality standards into executable checks for data and pipeline code; open-source, standardized transformation infrastructure leveraging best practices from the technology industry
 - *Example: Use of dbt, an open-source data transformation tool with native testing, documentation, and traceability features, to implement data transformation pipelines¹⁵*
- **Decision-making guidance** – a framework connecting pipeline design decisions to their impact on study validity
 - *Example: Extension of study diagnostics methods to assess sensitivity of study results to underlying data and pipeline quality*

The OHDSI community is well-positioned to benefit from and contribute to these innovations. The success of the OHDSI mission hinges on the reliability of the hundreds of OMOP CDMs in use around the world – but we currently lack a way to holistically verify those CDMs' quality. Fortunately, between institutional knowledge from OMOP sites, technical expertise from the OHDSI open-source community, and recent advances in AI technology, we have everything we need to operationalize this agenda and build a robust data foundation for evidence at scale.

*Regulatory agencies: European Health Data Space (EHDS), European Medicines Agency (EMA), United Kingdom Medicines & Healthcare Products Regulatory Agency (MHRA), United Kingdom National Health Service (NHS), United States Centers for Medicare and Medicaid Services (CMS), United States Food and Drug Administration (FDA), United States National Institutes of Health (NIH); NGOs: Duke Margolis Institute for Health Policy, Council for International Organizations of Medical Sciences (CIOMS), International Council for Harmonization (ICH), International Society for Pharmacoepidemiology (ISPE), International Society for Pharmaceutical Engineering (ISPE), International Society for Pharmacoeconomics and Outcomes Research (ISPOR), National Evaluation System for Health Technology (NEST), Transcelerate Biopharma; Data networks: Data Analysis and Real World Interrogation Network (DARWIN EU), European Health Data & Evidence Network (EHDEN), FDA Sentinel, Observational Health Data Sciences and Informatics (OHDSI), National Patient-Centered Clinical Research Network (PCORNet), PEDSNet.

**Citation list for literature review available upon request.

References

1. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Med Res Methodol.* 2022;22(1):287. doi:10.1186/s12874-022-01768-6
2. Priou S, Kempf E, Flicoteaux R, et al. "Where have my patients gone?": A simulation study on real-world data processing in Clinical Data Warehouses. *Health Policy Technol.* 2024;13(3):100893. doi:10.1016/j.hlpt.2024.100893
3. Lee K, Weiskopf N, Pathak J. A Framework for Data Quality Assessment in Clinical Research Datasets. *AMIA Annu Symp Proc.* 2017;2017:1080-1089.
4. Lee H, Xiong C, Baughman D, et al. A multidimensional hierarchical framework for sources of bias in real-world healthcare evidence: a scoping review. *J Biomed Inform.* Published online February 2026:104989. doi:10.1016/j.jbi.2026.104989
5. Gianfrancesco MA, Goldstein ND. A narrative review on the validity of electronic health record-based research in epidemiology. *BMC Med Res Methodol. BioMed Central Ltd.* 2021;21(1). doi:10.1186/s12874-021-01416-5
6. Arena PJ, Sun Y, Jaksa A, et al. Information Bias in Electronic Health Records and Administrative Claims Data: A Targeted Review of the Recent Literature. *Clin Pharmacol Ther. John Wiley and Sons Inc.* Preprint posted online February 1, 2025. doi:10.1002/cpt.70105
7. Wang S V., Sreedhara SK, Schneeweiss S, et al. Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions. *Nat Commun.* 2022;13(1):5126. doi:10.1038/s41467-022-32310-3
8. Kent S, Burn E, Dawoud D, et al. Common Problems, Common Data Model Solutions: Evidence Generation for Health Technology Assessment. *Pharmacoeconomics.* 2021;39(3):275-285. doi:10.1007/s40273-020-00981-9
9. Aten A, D'Ambrosio M, Hendricks-Sturup R. *Harnessing the Potential of Real-World Evidence Master Protocols.* Duke Margolis Center for Health Policy; 2023. Accessed May 6, 2026. <https://healthpolicy.duke.edu/publications/harnessing-potential-real-world-evidence-master-protocols>
10. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association.* 2012;19(1):54-60. doi:10.1136/amiajnl-2011-000376
11. Platt R, Carnahan RM, Brown JS, et al. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. *Pharmacoepidemiol Drug Saf.* 2012;21(S1):1-8. doi:10.1002/pds.2343
12. Rijnbeek PR. Converting to a Common Data Model: What is Lost in Translation? *Drug Saf.* 2014;37(11):893-896. doi:10.1007/s40264-014-0221-4
13. ChatGPT. Accessed April 16, 2025. <https://chat.openai.com/>
14. Observational Health Data Sciences and Informatics. DataQualityDashboard. Accessed October 8, 2026. <https://github.com/OHDSI/DataQualityDashboard>

15. dbt Labs. dbt Developer Hub. Accessed May 6, 2026. <https://docs.getdbt.com/?version=1.12>