Data Quality Framework for Assessing Data from Electronic Health Record-Enabled Registries

Amanda L. Lien, ND[1], Rishabh Jain, PhD[1], Claire A. Margolis, MS[1], Sam Haaf, MS[1], Alex Koshta, PhD[1], Solmaz Bagherpour, MS[1], David Budd, MSc[1], Srikanth Kaja, MS[1], Kevin Wood, MS[1], Andrew LaPrise, BS[1], Theodore Leng, MD[1], Matthew Roe, MD, MHS[1], Chhaya Shadra, MS[1]

[1]Verana Health, San Francisco, California, US

Introduction

Verana Health partners with leading medical associations to harness the comprehensive EHR data found in clinical data registries. We generate real-world evidence from these de-identified data to accelerate research and advance patient care.

Expanded uses for electronic health record (EHR) data, including scientific research and regulatory applications, have stimulated the need to delineate the fitness of EHR data across multiple dimensions. Rigorous data quality (DQ) assessment approaches are needed, but building broadly-applicable methods is challenging, compounded by the lack of clear standards to measure DQ[1]. General models often consider a limited set of quality dimensions, data sources, and use cases, and may not cover all aspects of DQ. Our holistic method applies layers of testing based on a matrix of quality dimensions (accuracy, completeness, consistency, generalizability, timeliness, traceability) across quality classes (technological, clinical, and scientific) that can be tailored to specific use cases.

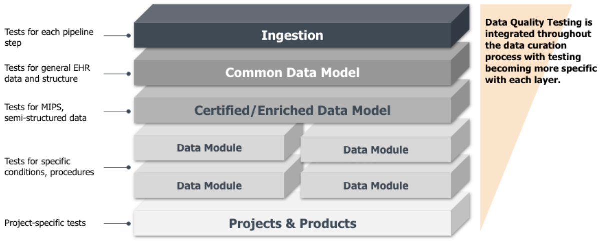
Cornerstones of Data Quality	Does the Data…	Technical	Clinical	Scientific
Completeness	*encompass the entire clinical picture?*	Field completeness is assessed among fields where data is expected. (e.g. each diagnosis must have a documented date)	Data completeness in clinical context exists (e.g., intraocular pressure is expected to be documented for patients with a diagnosis of Glaucoma)	Factors (e.g., confounders) have been considered in study design & analyses
Accuracy	*accurately reflect patient chart/reality?*	Data conforms to expected data types & constraints	EHR effectively captures patient journey & provider patterns	Results are within range of scientific acceptability
Traceability	*contain provenance back to source?*	Data elements & transformations are clear and auditable during ingestion and curation	Study specifies a clear, auditable patient cohort	Study design, methods, and analysis are clear and transparent
Consistency	*maintain integrity across structures, time, releases?*	Data are represented in a consistent data model, under congruent architecture & format	Cohort-specific trends & rates are tracked across time	Data is validated against published studies & external sources
Generalizability	*represent a minimally-biased sample?*	Data elements are harmonized to industry standards	Biases have been assessed & accounted for in clinical interpretation	External comparisons are used to identify and adjust for biases
Timeliness	*reflect recent practice patterns?*	Data is refreshed at appropriate frequency	Current practice patterns, treatments are incorporated	Data timeframe is relevant to current study

Table 1. Foundational concepts of Verana Health's data quality framework

Methods

We tested our framework on a market trend analysis utilizing data from the American Academy of Ophthalmology IRIS® Registry (Intelligent Research in Sight), the nation's first comprehensive eye disease clinical database. As of September 2020, data from 60 million patients and 15,000+ physicians using 60 different EHRs can be found in the database[2]. Tests were applied across quality dimensions and classes to a cohort of practices performing intravitreal anti-VEGF injections between 2018-Q2 and 2019-Q1. Our framework assesses the data quality of patients, providers, practices, and EHR based on a layer of a selected series of tests along the data curation pipeline. Each dimension score is comprised of the weighted average of multiple tests for that dimension and level.

Layers of Data Quality Tests


Tests for each pipeline step
Tests for general EHR data and structure
Tests for MIPS, semi-structured data
Tests for specific conditions, procedures
Project-specific tests
Ingestion
Common Data Model
Certified/Enriched Data Model
Data Module
Data Module
Data Module
Data Module
Projects & Products

Data Quality Testing is integrated throughout the data curation process with testing becoming more specific with each layer.

Figure 1. Data quality tests are applied at each step of the curation pipeline, as well as at the project and product levels.

Methods (continued)

For example, high level general tests such as checking for % nulls for each field are applied at the ingestion layer as an assessment for completeness of data. As our data becomes more curated in the certified/enriched layer, we apply more clinically specific tests such as % of patient diagnosis dates with corresponding procedure date for completeness of procedure records. As our data becomes further curated to condition-specific data models, we apply condition-specific tests such as % of anti-VEGF injections with retinal indication on the same day and eye. After applying multiple tests for each dimension, we aggregate these test scores into dimension scores for each practice. Each dimension score is then aggregated to provide an overall practice scores. Data from practices with an overall data quality score greater than the designated threshold are then used to run the market analysis project. We statistically compare the representativeness (gender, age, race, location) of the sub-cohort of high-quality patients to the original total cohort to make sure we did not create potential bias.

Results

The DQ framework identified practices with accurate, consistent, complete, traceable, and timely data related to intravitreal anti-VEGF use. Practices with low-quality data were then excluded to improve reliability of our analysis.

	Accuracy[a]	Completeness[b]	Consistency[b]	Timeliness[b]	Traceability[b]
All Practices (1048 practices)	72%	97%	86%	86%	98%
Reliable Practices Only (773 practices (74%))	100%	100%	100%	100%	100%
Less Reliable Practices Only (275 practices (26%))	9%	92%	59%	59%	83%

Table 2. a. based on a threshold of >= 0.6 on a scale of 0 to 1 (0.6 accounts for off-label use) b. based on a threshold of 1.0 on a scale of 0 to 1.0

Conclusions

Our framework proposes an approach to measuring and creating a more unified standard for assessing DQ. The framework, initially applied to the IRIS Registry, is designed to be applicable to other EHR-based registry data including the American Urological Association AQUA Registry and the American Academy of Neurology Axon Registry®, which use de-identified data for research purposes. Expanding our approach to pharmacovigilance and retrospective cohort studies, comparing across data refreshes, and assessing score reliability and sensitivity will be important to further validate model flexibility.

References

1. Cai L, Zhu Y. The challenges of data quality and data quality assessment in the big data era. Data Science Journal [Internet]. 2015 May 22;14:2–10. Available from: http://dx.doi.org/10.5334/dsj-2015-002
2. IRIS Registry. American Academy of Ophthalmology, 2020, www.aao.org/iris-registry.
3. Miksad, RA and Abernethy, AP. Harnessing the power of real-world evidence (RWE): A checklist to ensure regulatory-grade data quality. Clin Pharm & Ther 2017;103:202-205.