RWD96

Implementation of a Real-World Data Quality Framework in a Nationwide Oncology Electronic Health Record-Derived Database

Emily Castellanos MD MPH[1]; Brett Wittmershaus BSE[1]; Sheenu Chandwani MPH, PhD[1]
[1] Flatiron Health, New York, NY

Background

- Multiple real-world data (RWD) quality frameworks have been released identifying key dimensions of quality
- Challenges exist in applying these frameworks to scaled Electronic Health Record (EHR) data due to complexity of data sources and curation methods
- We demonstrate the application of a RWD quality framework incorporating core published dimensions of data quality to Flatiron Health RWD, a scaled, electronic health record (EHR)-based oncology RWD source

Methods

- We performed a targeted review of the following frameworks to identify dimensions of quality:
  - European Medicines Agency (Sep 2022)
  - National institute for Health and Care Excellence (Jun 2022)
  - United States Food and Drug Administration (Sep 2021)
  - Duke-Margolis Center for Health Policy (Aug 2019 and Oct 2018)
  - Patient-Centered Outcomes Research Institute (Sep 2016)
- We then reviewed data curation and quality assessment approaches for Flatiron Health RWD, curated and de-identified from longitudinal patient-level EHR-derived data generated during routine clinical practice originating from a nationwide network of US academic and community cancer practices (~3.4 million patient records)
- Quality processes were mapped to quality dimensions across published frameworks

Results

Figure 1: Source(s) of Data Variables in Flatiron Health RWD

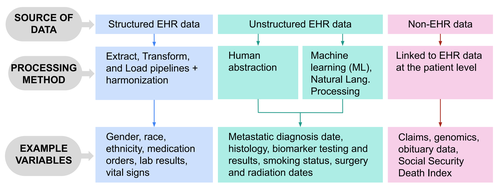
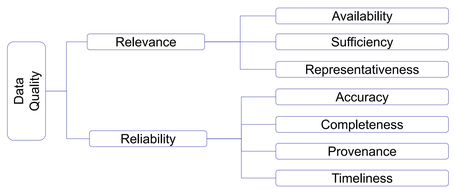


Table 1: Data Quality Dimensions in Flatiron Health RWD and Published Frameworks



Results cont'd

Relevance

- Defined as availability of critical variables and sufficient numbers of representative patients within the appropriate time period to address a given use case
- Relevancy to broad or specific use cases is optimized through dataset size, breadth, and depth of variables. Key features supporting relevancy include:
  - Core variables selected by oncology clinicians to support cohort inclusion criteria, exposures, outcomes, and covariates within a wide range of use cases
  - Broad representativeness in comparison to SEER and NPCR populations
  - Supplementation with deeper curation and data integrations for specific use case needs

Figure 2: Relevancy in Flatiron Health RWD

Reliability

- Defined as degree to which the data represent the clinical concept intended, inclusive of data accuracy, completeness, provenance, and timeliness

Accuracy:

- Defined as closeness of agreement between the measured value and the true value of what is intended to be measured
- Addressed with a **range of validation and verification approaches**, selected based on feasibility and criticality of the variable to the intended use case
  - Validation approaches: **external validation**, **internal validation**, and **indirect benchmarking**
  - Verification checks address **conformance**, **plausibility**, or **consistency**

Figure 4: External Validation of Composite Mortality Variable validated using the National Death Index

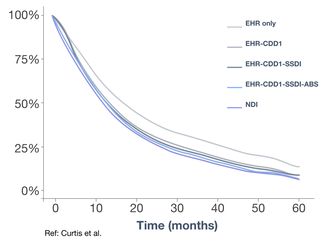


Figure 5: Indirect Benchmarking Validation of Novel Real World Progression Variable by Correlation to Related Endpoints

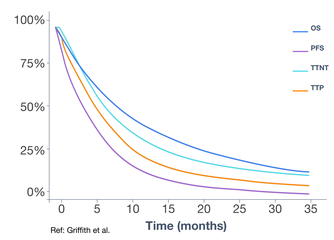


Figure 6: Internal Validation of ML-extraction vs Human Abstraction using a Replication Analysis

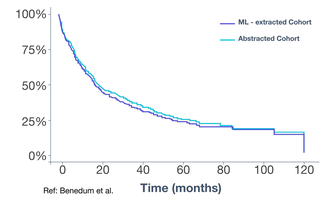


Figure 7: Verification check example

Temporal Plausibility: Treatment start date in close proximity to advanced diagnosis date

Completeness:

- Defined as presence of data values, without reference to actual values themselves
- Completeness approaches are intended to maximize the likelihood that information, if available within the source EHR, is included in the RWD

Figure 8: Examples of Approaches to Completeness across the Data Lifecycle

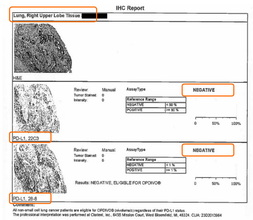
Structured EHR Data	Unstructured Human Abstracted Data	Unstructured ML/NLP Extracted Data
• Maximized by timeliness of data captured & integrity of data pipelines • Data refresh occurs with 24 hour recency • Sites with low completeness excluded • QC checks to detect pipeline/integration issues	• Logic checks within abstraction forms to minimize entry errors • Completeness distribution metrics evaluated prior to data freeze • Root cause investigations when less than expected completeness to inform abstractor guidance or flags for secondary review	• ML model developed to tune sensitivity of capture based on training data • Sensitivity metrics of data capture against validation data monitored in real-time • Completeness distributions expected to reflect EHR data availability

Provenance and Timeliness:

- Provenance accounts for the origin of a piece of data (in a database, document or repository) together with an explanation of how and why it got to the present place
- Timeliness addresses whether data are collected and curated with acceptable recency such that the dataset represents reality during the period of coverage

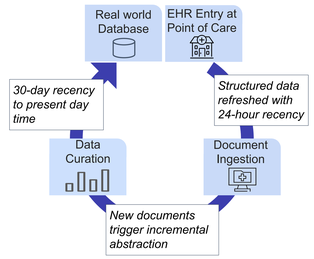
Figure 9: Provenance of Data Variables via Source Documentation Availability

Section of a PD-L1 Report, which is source documentation for selected data elements



- Tissue collection site
- Result
- Assay
- Lab Name

Figure 10: Timeliness in Data Pipeline Monitoring and Refresh Frequency



Discussion

- Addressing data quality in EHR-based RWD requires integration of systematic quality processes throughout the data lifecycle
  - Breadth of source data and access to it are foundational in enabling generation of high quality data
  - Knowledge of clinical data source, curation processes, clinical/scientific expertise, and use case needs are critical to optimize data quality processes
- Having a range of approaches allows optimization of quality processes to variable criticality, complexity, and need
- Limitations:
  - Flatiron Health RWD quality processes were aligned to dimensions in published frameworks; novel methods or sources of RWD may require new dimensions
  - RWE quality assessment also requires consideration of study design and analytic approach
  - Further research into setting standards for how quality should be assessed and communicated according to specific use cases is still needed

Conclusions

- Development of high quality, scaled EHR-based RWD requires integration of systematic processes across the data lifecycle
- Approaches to quality are optimized by knowledge of the clinical data source, curation processes, and use case needs
- By addressing quality dimensions from published frameworks, Flatiron Health RWD enables transparency in determining fitness for use

References

- Center for Drug Evaluation and Research Center for Biologics Evaluation and Research Oncology Center of Excellence. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products; Draft Guidance for Industry. US Food & Drug Administration Web site
- 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):1244-9214.1244. eCollection 2016. doi:10.13063/2327-9214.1244
- Griffith SD, Miksad RA, Calkins G, et al. Characterizing the feasibility and performance of real-world tumor progression end points and their association with overall survival in a large advanced non-small-cell lung cancer data set. *JCO Clin Cancer Inform.* 2019;3:1-13. doi:10.1200/CCI.19.00013
- Curtis MD, Griffith SD, Tucker M, Taylor MD, Capra WB, Carrigan G, Holzman B, Torres AZ, You P, Arnieri B, Abernethy AP. Development and Validation of a High-Quality Composite Real-World Mortality Endpoint. Health Serv Res. 2018 Dec;53(6):4460-4476.
- Benedum CM, Sondhi A, Fidyk E, Cohen AB, Nemeth S, Adamson B, Estévez M, Bozkurt S. Replication of Real-World Evidence in Oncology Using Electronic Health Record Data Extracted by Machine Learning. Cancers (Basel). 2023 Mar 20;15(6):1853.
- European Medicines Agency. Data Quality Framework for EU medicines regulation. 2022
- NICE real-world evidence framework: assessing data suitability. NICE Web site. https://www.nice.org.uk/corporate/ecd9/chapter/assessing-data-suitability#assessing-data-suitability. Updated 2022. Accessed March 20, 2023
- Duke-Margolis Center for Health Policy. Determining real-world data's fitness for use and the role of reliability. 2019
- Duke-Margolis Center for Health Policy. Characterizing RWD quality and relevancy for regulatory purposes. 2018

DISCLOSURES: All authors report employment at Flatiron Health, Inc., which is an independent member of the Roche Group, and stock ownership in Roche. EC and BW report equity ownership in Flatiron Health, Inc..