




6 Raising the Bar for Real-World Data in Oncology: Approaches to Quality Across Multiple Dimensions

Emily H. Castellanos, MD, MPH¹ ; Brett K. Wittmershaus, BSE¹ ; and Sheenu Chandwani, MPH, PhD¹ 

DOI <https://doi.org/10.1200/CCI.23.00046>

ABSTRACT

PURPOSE Electronic health record (EHR)–based real-world data (RWD) are integral to oncology research, and understanding fitness for use is critical for data users. Complexity of data sources and curation methods necessitate transparency into how quality is approached. We describe the application of data quality dimensions in curating EHR-derived oncology RWD.

METHODS A targeted review was conducted to summarize data quality dimensions in frameworks published by the European Medicines Agency, The National Institute for Healthcare and Excellence, US Food and Drug Administration, Duke–Margolis Center for Health Policy, and Patient–Centered Outcomes Research Institute. We then characterized quality processes applied to curation of Flatiron Health RWD, which originate from EHRs of a nationwide network of academic and community cancer clinics, across the summarized quality dimensions.

RESULTS The primary quality dimensions across frameworks were *relevance* (including subdimensions of availability, sufficiency, and representativeness) and *reliability* (including subdimensions of accuracy, completeness, provenance, and timeliness). Flatiron Health RWD quality processes were aligned to each dimension. Relevancy to broad or specific use cases is optimized through data set size and variable breadth and depth. Accuracy is addressed using validation approaches, such as comparison with external or internal reference standards or indirect benchmarking, and verification checks for conformance, consistency, and plausibility, selected on the basis of feasibility and criticality of the variable to the intended use case. Completeness is assessed against expected source documentation; provenance by recording data transformation, management procedures, and auditable metadata; and timeliness by setting refresh frequency to minimize data lags.

CONCLUSION Development of high-quality, scaled, EHR-based RWD requires integration of systematic processes across the data lifecycle. Approaches to quality are optimized through knowledge of data sources, curation processes, and use case needs. By addressing quality dimensions from published frameworks, Flatiron Health RWD enable transparency in determining fitness for real-world evidence generation.

ACCOMPANYING CONTENT

 Appendix

Accepted October 17, 2023

Published January 19, 2024

JCO Clin Cancer Inform

8:e2300046

© 2024 by American Society of
Clinical Oncology

Creative Commons Attribution
Non-Commercial No Derivatives
4.0 License

BACKGROUND AND OBJECTIVE

Real-world data (RWD) are data relating to patient health status and/or health care delivery collected from sources including electronic health records (EHRs), billing claims, and product and disease registries, while real-world evidence (RWE) provides conclusions resulting from RWD analysis.¹ The 21st Century Cures Act and increased adoption of health information technology have accelerated the growth of RWD sources and application of RWE to improve disease understanding, support the drug development lifecycle, and optimize patient care.²

The use of RWE is of great interest in oncology for many reasons, including the high unmet medical need, impact on quality of life, and urgency to improve patient outcomes.^{3–5} Oncology RWD rely heavily on EHR-based sources, which provide clinically meaningful longitudinal information about patients' characteristics, disease, and treatment course, enabling investigators to examine a wide range of use cases pertaining to diagnostic and therapeutic interventions.^{3,4} The urgency to harness the potential of EHR data to improve cancer care has been highlighted through the Cancer Moonshot⁶ and Childhood Cancer Data Initiatives,⁷ which support the collection of nationwide oncology RWD.

Curation of EHR-based RWD poses unique challenges that can affect quality. Documentation within the EHR exists in a wide range of formats, with much of the richest clinical information relevant for oncology research (eg, tumor histology or clinical outcomes) existing in unstructured documents requiring specialized curation processes.⁸ Although human abstraction has been a gold standard approach (unpublished data), it can be prohibitive in building timely and scalable data sets, an important consideration in oncology where the landscape is changing rapidly and large cohorts are often required to identify rare populations of interest.^{9,10} Advancements, such as machine learning (ML), have the potential to further unlock scale^{11,12} but require quality assessment to gain confidence and widespread adoption.

Fragmentation of health information is another challenge, with individual patient data often contained in different systems, that may or may not be designed for oncology workflows.¹³ This fragmentation poses challenges in interoperability¹⁴ and maintenance of data pipelines. Often, EHR source records are not fully accessible, limiting the implementation of standard and repeatable data quality standards.¹⁵ EHR data are frequently pooled from multiple sites and software programs to scale cohort sizes or integrated with non-EHR sources to increase completeness and/or identify comprehensive information relevant for use cases, requiring appropriate harmonization processes and validation techniques. Another challenge is identifying validation methods to demonstrate measurement accuracy.¹⁵ Although EHR-based RWD have been commonly used as a benchmark to validate other data sources (eg, claims), there is a frequent lack of external references more reliable than human abstraction to validate curated EHR data (unpublished data).¹⁶⁻¹⁹

Users of oncology RWD, including clinical and health services researchers, drug developers, health authorities, and regulatory bodies,²⁰ need to understand its quality to determine its fitness for evidence generation. This need is

manifested by the number of frameworks and regulatory guidances for assessing the fitness of RWD released globally in recent years.^{16,21-25} Although existing frameworks represent general considerations of data quality, to fully characterize the quality of RWD sources, it is important to establish transparency on how quality concepts are applied across the data curation lifecycle. Moreover, terminology and organizational approaches vary, and harmonization of concepts across frameworks is critical for comprehensively addressing RWD quality. In this paper, we describe the application of data quality dimensions in a scaled, EHR-based oncology RWD source developed to build fit-for-use data sets for secondary research.

DATA SOURCES

Flatiron Health RWD are curated from longitudinal patient-level EHR-derived data generated during routine clinical practice originating from a nationwide network of US academic and community cancer practices (approximately 3.4 million patient records).^{26,27} Data are ingested at source and curated into common data models.

Source data are classified as structured or unstructured (Fig 1). Structured data exist within the EHR in structured fields for the primary purpose of enabling clinic workflows and capturing patient-provider interactions. These data are extracted from the EHR across different sites and systems via secondary data processing and harmonized to computable and interoperable standard terminologies. EHR structured data variables include, but are not limited to, demographics (eg, birth year, sex, race/ethnicity, etc), vitals (eg, height, weight, etc), visits, laboratory data, practice information, diagnosis codes, medication orders, medication administrations, Eastern Cooperative Oncology Group (ECOG) performance status, and insurance coverage. Unstructured data include information such as clinic and nursing notes; laboratory, radiology, and pathology reports; and patient communications. Secondary data processing with human abstraction or technologies such as ML extraction and/or natural language processing (NLP) is

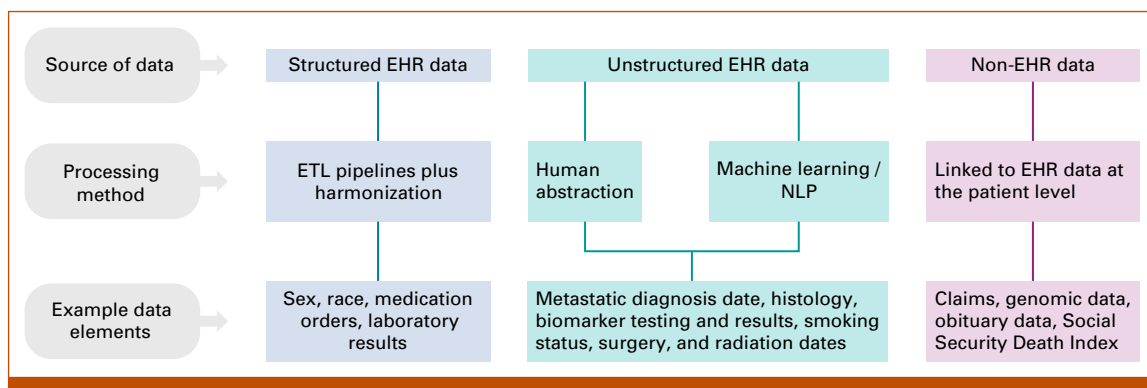


FIG 1. Source(s) of data variables in Flatiron Health real-world data. EHR, electronic health record; ETL, extract, transform, and load; NLP, natural language processing.

applied to curate unstructured data into structured variables.^{28,29} These EHR-derived data are linkable with non-EHR data sources, such as genomic or claims data to develop integrated clinicogenomic or clinicoclaims RWD.^{30,31}

ASSESSING RWD QUALITY

To establish the key data quality dimensions essential to RWD, we conducted a targeted review of frameworks and guidances on RWD quality published between 2015 and 2022 from five national or international health authorities and reimbursement agencies: The US Food and Drug Administration (FDA), European Medicines Agency (EMA), The National Institute for Healthcare and Excellence (NICE), the Pharmaceuticals and Medical Devices Agency, and the Canadian Agency for Drugs and Technologies in Health. We reviewed these five organizations for frameworks or guidances pertaining to RWD quality through agency websites and summaries of Really Simple Syndication (RSS) feeds, excluding publications irrelevant to EHR data, describing a single data source, or not pertaining to use of RWD for drugs and biologic decision making. From these five organizations, three publications were selected through discussion by a multidisciplinary team. Three additional publications from two nongovernmental organizations were included as primary sources, as these were frequently referenced by the above agency publications. Thus, the final review included six RWD quality frameworks and guidances from five organizations: EMA (September 2022), NICE (June 2022), FDA (September 2021), Duke-Margolis Center for Health Policy (August 2019 and October 2018), and the Patient-Centered Outcomes Research Institute (September 2016).^{16,21-25}

Data quality dimensions were independently reviewed; term definitions were extracted from each reference. Dimensions were included if described as distinct concepts in at least two frameworks; Appendix Table A1 lists dimensions excluded for inconsistency across sources or conceptual overlap with other dimensions. The hierarchy of dimensions and subdimensions was determined by the primary author based upon the organizational structure seen across different frameworks and independently reviewed by the two other authors; disagreements were resolved by group discussion.

Relevance and *reliability* were identified as the two primary quality dimensions with a set of subdimensions within each. Quality dimensions applied in generation of Flatiron Health RWD were then compared with the harmonized published dimensions summarized in Table 1. Methods applied in Flatiron Health RWD to assess each quality dimension are described below, including key examples to demonstrate alignment to published frameworks.

RELEVANCE

Relevance of RWD is defined as the availability of critical variables (exposure, outcomes, and covariates) and having a

sufficient number of representative patients within the appropriate time period to address a given use case. Flatiron Health has direct access to EHR systems that capture cancer patients' clinical workflows, which enhances the availability and clinical depth of oncology-relevant data. Data sets may be designed to address specific use cases or, in the case of scaled data sets, sets of potential use cases such as studies of natural history, unmet medical need, treatment effectiveness and safety, treatment patterns, prognostic or predictive biomarkers, or health policy.

Availability of Critical Variables

Development of Flatiron Health RWD begins by identifying critical variables for the final data model, which focuses quality efforts on the areas of highest importance to the use case. For multipurpose RWD, a set of core variables, including sociodemographics, cancer diagnosis, biomarkers, treatments, and clinical outcomes, are selected for inclusion by oncology clinicians to maximize breadth and ensure presence of commonly required inclusion/exclusion criteria, exposures, outcomes, and covariates. If core variables are insufficient for a specific use case, additional curation of variables from source EHR (eg, additional mutation subtypes, specific radiation types) or linkage to additional sources (eg, genomic data, claims data) is done to gather requisite data on the target population.

Sufficiency and Representativeness of Population

Sufficiency of population refers to whether the RWD provide an adequate number of relevant patients of interest to meaningfully address the intended use case(s). Flatiron Health RWD are curated with access to patient-level information to the time of EHR system initiation by the clinic, and most have patient records dated as far back as January 1, 2011, for capture of longitudinal clinical history. The large, nationwide patient network, variety of clinic sites, and coverage period enables sufficiency to address a wide range of oncology use cases. Representativeness describes the extent of similarity between sample and target populations, which may be broad or narrow depending upon the research question. Broad representativeness of Flatiron Health's RWD demographic and geographic distribution was established through comparison^{26,32} with the Surveillance, Epidemiology, and End Results Program and the National Program of Cancer Registries—both cardinal sources for population cancer surveillance and research in the United States.

RELIABILITY

Reliability is defined as the degree to which data represent the clinical concept intended and is assessed using four subdimensions: accuracy, completeness, provenance, and timeliness (Table 2). These subdimensions can be assessed independent of use case, as they primarily concern the ability of RWD to reflect reality.

TABLE 1. Data Quality Dimensions in Flatiron Health RWD in Comparison With Published Frameworks

Data Quality Dimension	Frameworks and Guidance	Definition
Relevance Availability, sufficiency, representativeness	Flatiron Health RWD	Availability of critical variables (exposure, outcomes, covariates) and sufficient numbers of representative patients within the appropriate time period to address a given use case
	EMA	Extent to which a data set presents data elements useful to answer a research question <i>Extensiveness, including coverage:</i> amount of information available with respect to what exists in the real world, whether it is within the capture process or not
	NICE	Determined by whether (1) the data provide sufficient information to produce robust and relevant results and (2) results are generalizable to patients in the NHS
	FDA	Availability of key data elements (exposure, outcomes, covariates) and sufficient numbers of representative patients for the study
	Duke-Margolis	Assessment of whether the data adequately address the applicable regulatory question or requirement, in part or in whole. Includes whether the data capture relevant information on exposures, outcomes, and covariates, and whether the data are generalizable
	PCORI	Contextual data quality features are described as entailing unique contextual or task-specific data quality requirements
Reliability	Flatiron Health RWD	Degree to which the data represent the clinical concept intended, inclusive of data accuracy, completeness, provenance, and timeliness
	EMA	The dimension that covers how closely the data reflect what they are designed to measure. It covers how correct and trustworthy the data are
	NICE	The ability to get the same or similar result each time a study is repeated with a different population or group
	FDA	Data accuracy, completeness, provenance, and traceability
	Duke-Margolis	Considers whether the data adequately represent the underlying medical concepts they are intended to represent; encompasses data accrual and data quality control (data assurance)
	PCORI	Intrinsic features of data values are described as features of quality that involve only the data values "in their own right" without reference to external requirements or tasks
Accuracy	Flatiron Health RWD	Closeness of agreement between the measured value and the true value of what is intended to be measured
	EMA	Amount of discrepancy between data and reality <i>Precision:</i> degree of approximation by which data represent reality
	NICE	How closely the data resemble reality
	FDA	Closeness of agreement between the measured value and the true value of what is intended to be measured <i>Validation:</i> the process of establishing that a method is sound or that data are correctly measured, usually according to a reference standard
	Duke-Margolis	Assessment of the validity, reliability, and robustness of a data field
	PCORI	Not defined; concepts of plausibility, conformance, and consistency are described as alternatives
Conformance	Flatiron Health RWD	Compliance of data values with internal relational, formatting, or computational definitions or internal or external standards
	EMA	Assesses coherence toward a specific reference or data model
	NICE	Whether the recording of data elements is consistent with the data source specifications
	FDA	Data congruence with standardized types, sizes, and formats
	Duke-Margolis	Congruence with standardized types, sizes, and formats; how compliant the data are with internal relational, formatting, or computational definitions or standards
	PCORI	Compliance of the representation of data against internal or external formatting, relational, or computational definitions. Data values align to specified standards and formats
Plausibility	Flatiron Health RWD	Believability or truthfulness of data values
	EMA	Likelihood of some information being true; a proxy to detect errors
	NICE	Not defined
	FDA	The believability or truthfulness of data values
	Duke-Margolis	Recorded values are logically believable given data source and expert opinion
	PCORI	Believability of data values (uniqueness, atemporal, temporal plausibility)
Consistency	Flatiron Health RWD	Stability of a data value within a data set or across linked data sets or over time
	EMA	Coherence: how different parts of overall data sets are consistent in their representation and meaning. Subdimensions include format coherence, structural coherence, semantic coherence, and uniqueness Uniqueness: same information is not duplicated but appears in the data set once
	NICE	Agreement in patient status in records across the data sources
	FDA	Included as part of the definition of data integrity: completeness, consistency, and accuracy of data
	Duke-Margolis	Stability of a data value within a data set or across linked data sets
	PCORI	Consistency is included as a subcategory of plausibility and conformance

(continued on following page)

TABLE 1. Data Quality Dimensions in Flatiron Health RWD in Comparison With Published Frameworks (continued)

Data Quality Dimension	Frameworks and Guidance	Definition
Completeness	Flatiron Health RWD	Presence of data values (data value frequencies, without reference to actual values themselves)
	EMA	Extensiveness, including completeness: amount of information available with respect to total information that could be available, given the capture process and data format
	NICE	Percentage of records without missing data at a given time point
	FDA	The “presence of the necessary data”
	Duke-Margolis	Measure of recorded data present within a defined data field and/or data set The frequencies of data attributes present in a data set without reference to data values
	PCORI	Frequencies of data attributes present in a data set, without reference to data values
Provenance	Flatiron Health RWD	An audit trail that 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
	EMA	Not defined
	NICE	Describes the ability to trace the origin of data and identify how it has been altered and transformed throughout its lifecycle. It provides an understanding of the trustworthiness or reliability of a data source
	FDA	An audit trail that “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” Traceability: permits an understanding of the relationships between the analysis results (tables, listings, and figures in the study report), analysis data sets, tabulation data sets, and source data
	Duke-Margolis	Origin of the data, sometimes including a chronologic record of data custodians and transformations Traceability: ability to record changes to location, ownership, and values Data accrual: the process by which data are collected and aggregated (includes provenance) Data lineage: the history of all data transformations (eg, recoding or modifying variables)
	PCORI	Not defined
Timeliness	Flatiron Health RWD	Data are collected and curated with acceptable recency such that the data set represents reality during the period of coverage
	EMA	Availability of data at the right time for regulatory decision making, that in turn entails that data are collected and made available within an acceptable time Currency: considers freshness of the data, eg, current and immediately useful Lateness: aspect of data being captured later than expected corresponding to reality
	NICE	Lag time between data collection and availability for research
	FDA	Not defined
	Duke-Margolis	Longitudinality: condition of data indexed by time/interval of exposure and outcome time
	PCORI	Not defined

NOTE. Duke-Margolis definitions are synthesized from both the August 2019 and October 2018 white papers.^{23,24}

Abbreviations: EMA, European Medicines Agency; FDA, US Food and Drug Administration; NHS, National Health Service; NICE, National Institute for Health and Care Excellence; PCORI, Patient-Centered Outcomes Research Institute; RWD, real-world data.

Accuracy

Accuracy is defined as the closeness of agreement between the measured value and true value of what is intended to be measured by the variable. Operational definitions for RWD variables are the codes or algorithms developed to curate source data and assign values.¹⁶ For curation of data from unstructured documents, Flatiron Health employs abstraction by trained clinical experts (eg, oncology nurses), using standardized policies as the operational definition. Because abstracted EHR data are often considered a reference standard for other RWD (eg, claims), a sufficiently available, high-quality external reference standard to validate abstracted variables is frequently lacking.

Accuracy: Validation Approaches

A range of validation approaches beyond external validation are applied to balance robustness and feasibility by

considering the risks and criticality of variables for an intended use (Fig 2). The robustness required is reflective of risk of misclassification of the variable, dictated by the likelihood of misclassification on the basis of variable complexity, and the potential consequence of misclassification to the use case on the basis of variable criticality.³⁴ Complex clinical concepts captured from unstructured data typically require a more robust validation process than variables that are structured data at the source.³⁵⁻³⁷ For example, variables such as birth year or sex are commonly entered into structured fields to support clinical workflows and are expected to be verified at multiple points during care. Although verification of data processing and pipelines remains essential to data integrity, validation of such data against external references is less critical, given that the value is expected to be accurate and complete to support clinical procedures. By assessing variable risk, resources can be efficiently deployed to optimize the overall quality of the scaled data set.

TABLE 2. Data Reliability Subdimensions in Flatiron Health RWD

Data Type	Structured Data	Unstructured Data	
Processing Method	Harmonization	Human Abstraction	ML/NLP Extraction
Accuracy	Data collected in structured format for primary purposes are harmonized to reference terminologies for secondary research use. Mapping processes are manually reviewed by the medical informatics team to ensure conformance to standards (external or Flatiron Health-established). Mapping guidelines are updated as new types of EHR data are available, or to reflect changes to secondary uses of EHR data	Data are validated using one or more of the following approaches: (1) Validation against an external reference standard; (2) Indirect benchmarking (data distributions, outcomes, etc) against literature and/or oncology clinical expert guidance; (3) Validation against an internal reference standard; and (4) Verification checks as proxies for accuracy	Data are validated using an internal reference standard, typically human abstracted data. Using the internal reference standard, metrics (eg, sensitivity, specificity, PPV, and NPV) are assessed
Completeness	Data completeness is reflective of data availability within the EHR and is maximized by ensuring timeliness of data capture and integrity of data pipelines. Sites with low completeness during integration are excluded until they meet target thresholds. Quality control checks detect any large drops in data that would signal issues with integrations or ETL pipelines	Data abstraction forms are built with logic checks to ensure data are input when needed. Data completeness distributions are assessed according to expectations for data availability within the EHR; if expectations are not met, then further investigation is conducted to find and correct the root cause	Data completeness is expected to reflect data availability within the EHR and is assessed by determining sensitivity of data capture against the reference data upon which the ML algorithm is trained and validated
Provenance	Data are traceable to the source. Harmonization rules dictating data mapping are maintained, updated as source data changes, and available as needed. Reference terminologies to which source data are mapped are updated, versioned, and maintained	Individual patient data points are traceable via a proprietary technology platform ^a with an audit trail of abstracted data inputs, changes, and source documentation from the EHR reviewed by trained clinical abstractors. Policies and procedures and data abstraction forms are version-controlled	Data are traceable to source documentation via audit trails for NLP-acquired text. The ML algorithm is archived, and algorithm updates are logged
Timeliness	Mapped data are refreshed on a 24-hour cadence. Data pipelines are continually monitored, and sites with stale structured data are excluded	New EHR documentation considered relevant for a given variable is identified and surfaced for abstraction with set recency (typically 30 days) to facilitate incremental updates. Abstraction resulting from new EHR documentation is reviewed and completed before data cutoff. Document ingestion is monitored	Information that is documented within the EHR by time of data cutoff, whether it exists in structured or unstructured formats, is ingested and processed such that it is available for ML extraction

Abbreviations: EHR, electronic health records; ETL, extract, transform, and load; ML, machine learning; NLP, natural language processing; NPV, negative predictive value; PPV, positive predictive value; RWD, real-world data.

^aShklarski et al.³³

Comparison of data values to an *external reference standard* is a preferred validation approach for an RWD variable when feasible. Ideally this is a gold standard containing true values of the desired clinical concept; however, these rarely exist. Instead, external reference standards, commonly accepted and presumably of superior quality, are used when available (unpublished data). For example, Flatiron Health RWD validated a composite mortality variable curated from different data sources (EHR data, the Social Security Death Index, and commercial obituary data) using the National Death Index as a reference standard.³⁶ In another case, Flatiron Health created an external reference standard of radiologist-measured response via RECIST 1.1 to validate clinician-assessed response on the basis of EHR documentation.³⁵ Although robust, such validations are often labor-intensive or infeasible.

When external validation is infeasible, Flatiron Health uses a range of prespecified and data-driven approaches to assess evidence of variable accuracy. One such approach is *indirect benchmarking* using the distribution of variable values and, where applicable, correlation with other data points in accordance with available literature and clinical expectations. For example, validation of a novel real-world progression variable included assessment of whether corresponding end points, such as real-world progression-free survival, showed expected correlations with temporal events such as death or treatment changes.³⁷

Another approach is to generate an *internal reference standard* from the same source data using a previously established curation method. This approach is most useful to validate the execution of an operational definition between two different data curation methods (eg, human abstraction and ML

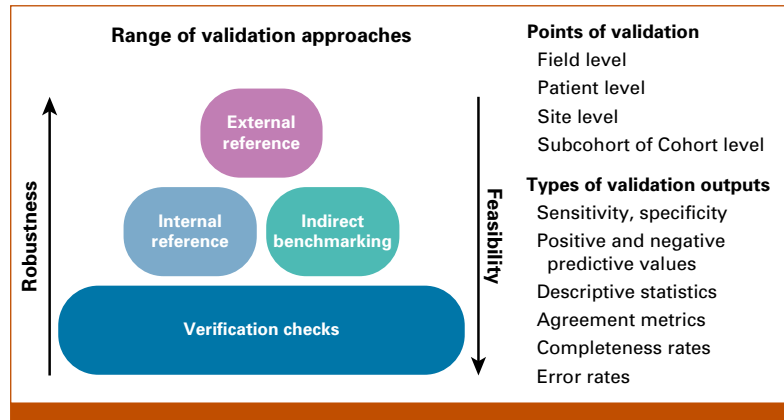


FIG 2. Approaches to accuracy in Flatiron Health real-world data.

extraction),³⁸ data sources (eg, unstructured and structured), or when transposing an operational definition developed in one application to another. For example, Flatiron Health creates internal reference standards using EHR abstraction by trained clinical experts to validate scaled algorithmic approaches, such as NLP or NLP-based ML-extracted data. In such cases, ML models are trained, validated, and assessed for bias using separate abstracted databases,²⁹ then monitored using abstracted data to ensure validation metrics (sensitivity, specificity, and positive predictive value) remain stable over time. This approach relies on the operational definition used to curate the internal reference standard reflecting the intended reality. Using an internal reference standard in combination with indirect benchmarking of established clinical knowledge can be a powerful combination to assess the measurement error of the definition and application of the operational definition. When these validation approaches are infeasible, robust verification processes become more important.

Accuracy: Verification Approaches

Verification checks play an important role in controlling and improving data quality, and can serve as a proxy for accuracy.²⁵ Clinical expertise, such as understanding how variables should logically relate to each other, is leveraged by consultation with oncology clinicians to address the following categories: *conformance*, compliance of data values with internal relational, formatting, or computational definitions or internal or external standards; *plausibility*, believability or truthfulness of data values; and *consistency*, stability of a data value within a database, across linked databases, over time, or between data processing approaches.

Table 3 describes examples of verification checks. In Flatiron Health RWD, verification checks are implemented across the data lifecycle: at patient and cohort levels, during data curation, and after data set construction. For example, abstraction processes include conformance, consistency, and

plausibility verification checks to prevent errors during entry. Random samples of data are duplicately abstracted, producing agreement measures to monitor consistency to a prespecified threshold, which alert to errors or shifts in EHR documentation over time. Once the data set is curated, cohort-level verification checks are implemented and reviewed. Investigation into the root cause(s) of errors found during verification checks informs continuous improvement efforts such as further guidance and training of abstractors or implementation of additional verification checks. Verification efforts are focused on critical variables or areas of highest complexity for the intended use case.

Completeness

Completeness is defined as the presence of data values, without reference to actual values themselves. At the field level, an observation's presence or absence in RWD is viewed as the combination of (1) whether the information is available in the source data (ie, recorded in the EHR) and (2) whether that information is entered into the RWD. An observation's value may be missing for a variety of reasons ranging from random chance (eg, data entry error) to intentional omission (eg, only recording abnormal results).

Flatiron Health's approach to completeness is intended to maximize the likelihood that information, if available within the source EHR, is included in the RWD. External references to benchmark RWD completeness are frequently lacking; thus, completeness of critical variables is commonly assessed according to clinical expectations for source documentation, standard site documentation practices, and/or internal reference standards (Appendix Table A2).²⁹ For example, a diagnostic laboratory may be clinically indicated for certain patients; as a result, data are expected to be more complete within the relevant subgroup relative to the full database. Sites that do not meet thresholds for structured data completeness are ineligible for inclusion. For unstructured data, completeness is assessed across the cohort and may include evaluations at any time point or within specific time windows. If data are less complete than

TABLE 3. Sample of Verification Checks in Flatiron Health RWD

Category	Subcategory	Description	Example Verification Check
Conformance	Value conformance	Data values conform to internal formatting constraints	Dates are recorded as YYYY-MM-DD
		Data values conform to allowable values or ranges	Stage is abstracted from unstructured documents into structured categories aligned to AJCC terminology
	Relational conformance	Data values conform to relational constraints	Patients with documentation of real-world response events also have documented treatment data
		Unique (key) data values are not duplicated	Duplicate records for the same patient across multiple clinic sites are merged into a single record
Computational conformance	Changes to the data model or data model versioning	Changes to the data model are tracked and inputs only allowed that match the current data model at the time of entry	
Plausibility	Uniqueness plausibility	Computed values conform to computational or programming specifications	Human-abstracted group stage and group stage calculated from abstracted T, N, M components, when available, are identical
		Data values that identify a single object are not duplicated	Biomarker tests are not captured in duplicate when there are multiple references to the same event in documentation
	Atemporal plausibility	Data values and distributions agree with an internal measurement or local knowledge (overlaps with indirect benchmarking)	First-line treatment regimens, as defined according to line of therapy business rules, reflect expected clinical practice as described by NCCN guidelines
		Data values and distributions for independent measurements of the same or related facts agree	Date of treatment discontinuation for disease progression is in close proximity to date of progression documented on imaging
		Logical constraints between values agree with local or common knowledge (includes "expected" missingness)	Patients receiving TRK inhibitor therapy have documentation of an <i>NTRK</i> fusion
		Biologic plausibility of different values is in agreement with local or common knowledge	Coexistence of <i>EGFR</i> and <i>KRAS</i> mutations are rare
	Temporal plausibility	Values of repeated measurement of the same fact show expected variability	Time between repeated response assessments is generally aligned to intervals recommended by NCCN guidelines; however, shorter and longer intervals are also present in line with real-world practice patterns
		Observed or derived values conform to expected temporal properties	Initial diagnosis date precedes metastatic diagnosis date for patients whose cancer stage at initial diagnosis is nonadvanced
		Sequences of values that represent state transitions conform to expected properties	Real-world progression events are followed by a logical clinical event, such as change in treatment, referral to hospice, or death
		Measures of data value density against a time-oriented denominator are expected on the basis of internal or common knowledge	PD-L1 testing events become more frequent after approval of a therapy, for which PD-L1 positivity is required by indication
Consistency	Cross-field consistency	Data are consistent across multiple fields or data sources	Patients documented as having brain metastases at initial diagnosis are also identified as having stage IV disease
		Data from recurring or refreshed databases are consistent over time	Frequency of PSA values within a given site shows minimal month-over-month variation
	Agreement	Duplicate capture of the same data point by different processes or individuals results in the same values	Two abstractors agree on the discontinuation date and reason for discontinuation of the same drug
	Reproducibility	Repeat use of operational data capture algorithms will result in the same or similar results	Performance of a smoking status variable leads to a consistent extracted result each time it is used on the same or similar tasks

NOTE. Modified from Kahn et al.²⁵

Abbreviations: AJCC, American Joint Committee on Cancer; *EGFR*, epidermal growth factor receptor; NCCN, National Comprehensive Cancer Network; *NTRK*, neurotrophic tyrosine receptor kinase; PD-L1, programmed death ligand 1; PSA, prostate-specific antigen; TRK, tropomyosin receptor kinase.

expected, causes of missingness are investigated and may lead to updates in abstractor guidance, improvements to ML models, or creation of quality controls to improve sensitivity of data collection. In certain cases, data from multiple sources are integrated to improve variable completeness. For example, ECOG performance status may be documented in structured or unstructured formats, and

completeness is optimized by combining both sources into a composite variable.³⁹

Provenance

Provenance accounts for the origin of a piece of data together with an explanation of how and why it got to the

present place. This includes each step an RWD variable goes through such that data transformations during processing, data management procedures, and auditable metadata are available. Flatiron Health has access to the EHR through data extracts or the actual EHR itself. Standard quality procedures are in place to ensure repeatable processes are followed, such as how duplicate records, data linking, and data cleaning are handled; example procedures are described in Appendix [Table A2](#).

For structured data, extract, transform, and load (ETL) pipelines are traceable through a controlled codebase to understand how source data are processed. Structured data mapping and harmonization steps are logged and have controlled rules. For unstructured data, Flatiron Health has access to free text from the EHR and/or the scanned documents used to create its databases. ETL pipelines for processing documents, abstraction, and ML data variables are tracked. A proprietary abstraction platform³³ logs available and viewed documents, abstraction policy version, abstraction certifications, notes, and data cleaning steps; the platform links each curated data point to its source documentation within the EHR. When ML is used to process unstructured data, the model version is logged providing provenance to the algorithm as well as model training and testing sets. When NLP is used to extract text and/or inform model outputs, Flatiron Health is able to identify the specific unstructured text snippets that provided the data.

Timeliness

Timeliness reflects whether RWD are curated with acceptable recency, such that it represents reality during the coverage period. When considering fitness for use, the concept of timeliness may be closely related to relevance,²⁵ as the coverage period of interest varies depending upon the research question. Although included as its own dimension in some frameworks but not all,^{21,40,41} Flatiron Health considers timeliness to be a core distinct attribute of RWD quality. The coverage period is regularly refreshed to align as closely as possible to present-day time. Structured data from the EHR are refreshed on a 24-hour cadence. Unstructured documents for review are similarly made available and curated at a predetermined frequency, most commonly a 30-day cadence. Data pipelines from each site are continuously monitored (see Appendix [Table A2](#) for examples), and sites that have stale data (ie, no new information within a recent time period) are excluded until issues are resolved.

Monitoring processes ensure timeliness of data capture with respect to time from primary data availability to inclusion within the RWD. However, because of the nature of RWD, variation in primary data availability is expected depending upon the clinical setting. For example, patients who are undergoing surveillance in the adjuvant setting may visit the clinic several times per year, whereas patients undergoing active treatment may be seen multiple times within a month.

Such real-world variation should be accounted for in study analyses.

DISCUSSION

In this paper, we demonstrate that Flatiron Health EHR-derived RWD are curated to be relevant and reliable using a range of quality processes that optimize for robustness, scalability, and feasibility. As established across published frameworks, quality must be assessed across multiple dimensions and all dimensions addressed within the data generation lifecycle. The breadth of source data and level of access to it are a critical foundation for generating high-quality data. With access to approximately 3.4 million EHR records, nationwide cancer population coverage, and over a decade-long availability of longitudinal clinical information, we are able to generate fit-for-use RWD to address a wide range of oncology use cases.

Having a range of data quality approaches that can be calibrated to different breadth and depth levels is critical to scalability. Data quality approaches are optimized when the most labor-intensive and scientifically rigorous processes are deployed for the most critical and complex data. Multidisciplinary expertise is leveraged to determine how to most efficiently apply different quality processes. Knowledge of medical informatics, including clinical data entry and flow of information within the EHR, is used to set appropriate expectations for completeness and calibrate validation methods to risk of misclassification. Clinical and scientific expertise is used to design validation studies, identify appropriate reference standards, and design appropriate statistical tests for evaluation. The robustness of the approach applied also considers the context in which data are expected to be used. Data intended for regulatory submission, for example, often require more robust validation and a deeper data model than commercial tracking.

Our approaches to RWD quality largely reside downstream of primary data collection, which minimizes implications on the entry of source EHR data itself. However, improving the quality of data at source is another avenue to scaling EHR data curation. For example, the Minimal Common Oncology Data Elements (mCODE) initiative was launched by ASCO to develop a consensus data standard for oncology.⁴² The utility of the mCODE standard is being further informed by the CodeX HL7 FHIR Accelerator, a diverse group of public and private stakeholders who are collaborating on projects to standardize oncology EHR data collection.⁴³ Because such efforts have the potential to affect clinical workflows, success of these initiatives should be measured by their effect on clinical documentation burden as well as RWD quality.

Lack of consistent terminology has been identified as a limitation when assessing quality across different RWD sources.²⁵ In our review, however, we found relative consistency in concepts that were considered critical to fitness for use, although some variation in terminology was seen. In

particular, conceptual overlap was found among the terms quality and reliability^{23,24}; in such cases where terminology evolved over time,^{23,24} we used the most recent terminology. Our final approach includes both relevance and reliability as part of data quality to capture as comprehensively as possible the dimensions that must be addressed to determine fitness for use.

Our application of a data quality framework to Flatiron Health's processes has several limitations. First, the field of RWD is evolving, and it is possible that novel sources, such as imaging or digital pathology, will require new dimensions to fully characterize their quality.²⁵ We intentionally limited our review to frameworks focused on RWD. Publications focusing on specific techniques, such as ML,⁴⁴ were excluded for this generalized RWD quality framework, as were evaluations of study design and analyses, although these should be considered when determining RWE fitness for use.⁴⁵ Finally, this paper broadly describes how quality approaches can be implemented but is not setting specific thresholds for quality as this depends upon

the context and purpose for which data are used. Further research addressing standards for how quality should be assessed and communicated according to specific use cases is still needed.

Transparency of RWD quality is key to unlocking their value to accelerate research and expand use across the drug development and care delivery landscape. Building scaled RWD from EHR sources, in which all dimensions of quality are addressed, is a complex endeavor requiring multi-disciplinary expertise across clinical medicine, data science and operations, engineering, and medical informatics. Investment in robust processes to curate high-quality RWD is a critical step to generating confidence in the use of the resulting RWE. By demonstrating how dimensions of data quality can be addressed within large-scale, EHR-derived oncology RWD, we show how clinically informed processes, assessments, and scientific methods are used to curate fit-for-use RWD and lay the groundwork for how novel approaches to collecting RWD from new or existing data sources are assessed for quality.

AFFILIATION

¹Flatiron Health, Inc, New York, NY

CORRESPONDING AUTHOR

Emily H. Castellanos, MD, MPH, Flatiron Health, Inc, 233 Spring St, New York, NY 10013; e-mail: ecastellanos@flatiron.com.

PRIOR PRESENTATION

Presented in part at the International Society for Pharmacoeconomics and Outcomes Research, Boston, MA, May 7-10, 2023.

SUPPORT

Supported by Flatiron Health, Inc, which is an independent member of the Roche Group.

AUTHOR CONTRIBUTIONS

Conception and design: All authors

Collection and assembly of data: Brett K. Wittmershaus, Sheenu Chandwani

Data analysis and interpretation: All authors

Manuscript writing: All authors

Final approval of manuscript: All authors

Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted.

I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians ([Open Payments](https://www.openpayments.gov/)).

Emily H. Castellanos

Employment: Flatiron Health

Stock and Other Ownership Interests: Flatiron Health, Roche

Brett K. Wittmershaus

Employment: Flatiron Health

Stock and Other Ownership Interests: Roche, Flatiron Health

Sheenu Chandwani

Employment: Merck, Flatiron Health

Stock and Other Ownership Interests: Merck, Flatiron Health

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

The authors thank Hannah Gilham, Darren Johnson, and Jennifer Swanson from Flatiron Health, Inc, for editorial support.

REFERENCES

- Sherman RE, Anderson SA, Dal Pan GJ, et al: Real-world evidence - what is it and what can it tell us? *N Engl J Med* 375:2293-2297, 2016
- Health Information Technology for Economic and Clinical Health (HITECH) Act, Title XIII of Division A and Title IV of Division B of the American Recovery and Reinvestment Act of 2009 (ARRA). Pub. L. no. 111-5, 123 Stat. 226 (Feb. 17, 2009) (Full-Text), Codified at 42 U.S.C. §§300jj Et Seq.; §§17901 Et Seq. <https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/hitechact.pdf>
- Visvanathan K, Levit LA, Raghavan D, et al: Untapped potential of observational research to inform clinical decision making: American society of clinical oncology research statement. *J Clin Oncol* 35:1845-1854, 2017
- Bolislis WR, Fay M, Kühler TC: Use of real-world data for new drug applications and line extensions. *Clin Ther* 42:926-938, 2020
- Feinberg BA, Gajra A, Zettler ME, et al: Use of real-world evidence to support FDA approval of oncology drugs. *Value Health* 23:1358-1365, 2020

6. US Government: The National Cancer Plan, 2023. <https://nationalcancerplan.cancer.gov/>
7. Flores-Toro JA, Jagu S, Armstrong GT, et al: The childhood cancer data initiative: Using the power of data to learn from and improve outcomes for every child and young adult with pediatric cancer. *J Clin Oncol* 41:4045-4053, 2023
8. Zhao Y, Howard R, Amorrtoru RP, et al: Assessing the contribution of scanned outside documents to the completeness of real-world data abstraction. *JCO Clin Cancer Inform* 7:e2200118, 2023 [10.1200/CCI.22.00118](https://doi.org/10.1200/CCI.22.00118)
9. Miksad RA, Samant MK, Sarkar S, et al: Small but mighty: The use of real-world evidence to inform precision medicine. *Clin Pharmacol Ther* 106:87-90, 2019
10. Liu J, Barrett JS, Leonardi ET, et al: Natural history and real-world data in rare diseases: Applications, limitations, and future perspectives. *J Clin Pharmacol* 62:S38-S55, 2022 (suppl 2)
11. Nagy M, Radakovich N, Nazha A: Machine learning in oncology: What should clinicians know? *JCO Clin Cancer Inform* [10.1200/CCI.20.00049](https://doi.org/10.1200/CCI.20.00049)
12. Swanson K, Wu E, Zhang A, et al: From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. *Cell* 186:1772-1791, 2023
13. Post AR, Burningham Z, Halwani AS: Electronic health record data in cancer learning health systems: Challenges and opportunities. *JCO Clin Cancer Inform* [10.1200/CCI.21.00158](https://doi.org/10.1200/CCI.21.00158)
14. Patt D, Stella P, Bosserman L: Clinical challenges and opportunities with current electronic health records: Practicing oncologists' perspective. *JCO Oncol Pract* 14:577-579, 2018
15. Cowie MR, Blomster JI, Curtis LH, et al: Electronic health records to facilitate clinical research. *Clin Res Cardiol* 106:1-9, 2017
16. 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. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory>
17. Alam AU, Karkhaneh M, Wu C, et al: Development and validation of a case definition to identify hemophilia in administrative data. *Thromb Res* 204:16-21, 2021
18. Gillmeyer KR, Nunez ER, Rinne ST, et al: Development and validation of algorithms to identify pulmonary arterial hypertension in administrative data. *Chest* 159:1986-1994, 2021
19. Esposito DB, Banerjee G, Yin R, et al: Development and validation of an algorithm to identify endometrial adenocarcinoma in US administrative claims data. *J Cancer Epidemiol* 2019:1938952, 2019
20. Basch E, Schrag D: The evolving uses of "real-world" data. *JAMA* 321:1359-1360, 2019
21. European Medicines Agency: Data Quality Framework for EU Medicines Regulation, 2022. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/data-quality-framework-eu-medicines-regulation_en.pdf
22. NICE real-world evidence framework: Assessing data suitability, June 23, update. <https://www.nice.org.uk/corporate/ecd9/chapter/assessing-data-suitability#assessing-data-suitability>
23. Duke-Margolis Center for Health Policy: Determining Real-World Data's Fitness for Use and the Role of Reliability, 2019. <https://healthpolicy.duke.edu/publications/determining-real-world-datas-fitness-use-and-role-reliability>
24. Duke-Margolis Center for Health Policy: Characterizing RWD Quality and Relevancy for Regulatory Purposes, 2018. <https://healthpolicy.duke.edu/publications/characterizing-rwd-quality-and-relevancy-regulatory-purposes-0>
25. 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)* 4:1244, 2016
26. Ma X, Long L, Moon S, et al: Comparison of population characteristics in real-world clinical oncology databases in the US: Flatiron health, SEER, and NPCR. *medRxiv* [10.1101/2020.03.16.20037143](https://doi.org/10.1101/2020.03.16.20037143)
27. Birnbaum B, Nussbaum N, Seidl-Rathkopf K, et al: Model-assisted cohort selection with bias analysis for generating large-scale cohorts from the EHR for oncology research. *arXiv* [10.48550/arXiv.2001.09765](https://arxiv.org/abs/2001.09765)
28. Adamson B, Washom M, Blarre A, et al: Approach to machine learning for extraction of real-world data variables from electronic health records. *medRxiv* [10.48550/arXiv.2001.09765](https://doi.org/10.48550/arXiv.2001.09765)
29. Estevez M, Benedum CM, Jiang C, et al: Considerations for the use of machine learning extracted real-world data to support evidence generation: A research-centric evaluation framework. *Cancers (Basel)* 14:3063, 2022
30. Agarwala V, Khozin S, Singal G, et al: Real-world evidence in support of precision medicine: Clinico-genomic cancer data as a case study. *Health Aff (Millwood)* 37:765-772, 2018
31. Ma X, Yerram P, Wang J, et al: Characterization of a novel oncology electronic health records-derived data linkage to commercial claims. Presented at the 38th International Conference on Pharmacoepidemiology and Therapeutic Risk Management (ICPE), Copenhagen, Denmark, August 26-28, 2022
32. Snow T, Snider J, Comment L, et al: Comparison of population characteristics in real-world clinical oncology databases in the US: Flatiron health-foundation medicine clinico-genomic databases, flatiron health research databases, and the national cancer institute SEER population-based cancer registry. *medRxiv* [10.1101/2023.01.03.22283682](https://doi.org/10.1101/2023.01.03.22283682)
33. Shklarski G, Abernethy A, Birnbaum B, et al, inventors; Flatiron Health, Inc., assignee: Extracting Facts From Unstructured Data. U.S. Patent Pending. Application No. US17/018,166. September 11, 2020
34. ISO 31000 risk management. <https://www.iso.org/iso-31000-risk-management.html>
35. Ma X, Bellomo L, Hooley I, et al: Concordance of clinician-documented and imaging response in patients with stage IV non-small cell lung cancer treated with first-line therapy. *JAMA Netw Open* 5:e229655, 2022
36. Zhang Q, Gossai A, Monroe S, et al: Validation analysis of a composite real-world mortality endpoint for patients with cancer in the United States. *Health Serv Res* 56:1281-1287, 2021
37. 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* [10.1200/CCI.19.00013](https://doi.org/10.1200/CCI.19.00013)
38. Washom ML, Tan K, Wiberg H, et al: A hybrid approach to scalable real-world data curation by machine learning and human experts. *medRxiv* [10.1101/2023.03.06.23286770](https://doi.org/10.1101/2023.03.06.23286770)
39. Cohen AB, Rosic A, Harrison K, et al: A natural language processing algorithm to improve completeness of ECOG performance status in real-world data. *Appl Sci* 13:6209, 2023
40. Liaw ST, Rahimi A, Ray P, et al: Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature. *Int J Med Inform* 82:10-24, 2013
41. Kahn MG, Raebel MA, Glanz JM, et al: A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Med Care* 50:S21-S29, 2012 (suppl)
42. Osterman TJ, Terry M, Miller RS: Improving cancer data interoperability: The promise of the minimal common oncology data elements (mCODE) initiative. *JCO Clin Cancer Inform* [10.1200/CCI.20.00059](https://doi.org/10.1200/CCI.20.00059)
43. CodeX HL7 FHIR accelerator, April 17, update. <https://confluence.hl7.org/display/COD/CodeX+Home>
44. Blueprint for trustworthy AI implementation guidance and assurance for healthcare, December 7, update. <https://www.coalitionforhealthai.org/insights>
45. Cocoros NM, Arlett P, Dreyer NA, et al: The certainty framework for assessing real-world data in studies of medical product safety and effectiveness. *Clin Pharmacol Ther* 109:1189-1196, 2021

APPENDIX

TABLE A1. Data Quality Concepts Not Included Within the Flatiron Health Framework (Table 1) Because of Inconsistent Use Across Sources or Overlap With Other Dimensions

Concept	Definition	Rationale for Excluding
Coherence	Dimension that expresses how different parts of an overall data set are consistent in their representation and meaning closely relates to consistency and validation. We can consider consistency and coherence largely synonyms, with the caveat that detection of inconsistencies is often a way to measure the reliability of data. Subdimensions of coherence include format coherence, structural coherence, semantic coherence, and uniqueness. Conformance assesses coherence toward a specific reference or data model	Coherence is only described as a distinct dimension by EMA Coherence overlaps with both consistency and conformance, which are incorporated into the evaluation of accuracy under the Flatiron Health framework
Coverage	Amount of information available with respect to what exists in the real world, whether it is within the capture process or not. This cannot be easily measured if the total information is not definable or accessible	Coverage is only described as a distinct dimension by EMA Flatiron Health's evaluative approach to completeness includes elements of coverage, as thresholds of completeness are set based on clinically informed expectations of information availability
Extensiveness	How much data are available, and whether the data are sufficient for purpose. Extensiveness is composed of completeness and coverage	Extensiveness is only described as a distinct dimension by EMA Extensiveness overlaps with dimensions of completeness and sufficiency in the Flatiron Health framework
Precision	Degree of approximation by which data represent reality. For instance, the age of a person could be reported in years or months	Precision is only described as a distinct dimension by EMA Precision or granularity of a variable is incorporated in part in accuracy, with the degree to which the operational definitions represents reality, and in part under the availability subdimension of relevance, with the variable having the appropriate granularity for the use case Excluded also to avoid the term precision, which can also be used to mean positive predictive value
Traceability	Permits an understanding of the relationships between the analysis results (tables, listings, and figures in the study report), analysis data sets, tabulation data sets, and source data	Traceability is only described as a distinct dimension by FDA Excluded from Flatiron Health data quality process as the focus of this dimension on analytic output is more applicable to real-world evidence than real-world data Traceability of source data is incorporated under provenance

Abbreviations: EMA, European Medicines Agency; FDA, US Food and Drug Administration; ML, machine learning.

TABLE A2. Sample of Quality Processes for Completeness, Provenance, and Timeliness in Flatiron Health Real-World Data

Dimension	Subcategory	Description	Example Quality Process
Completeness	Site level	Site-level completeness is assessed across selected variables, with thresholds set based on variable criticality and clinical or internal benchmarks	Completeness thresholds for critical laboratory data range from 40% (eg, lactate dehydrogenase) to 90% (eg, hematocrit and hemoglobin) Completeness targets on the basis of median site scores (eg, Route of medication administration documentation rate is >70%) Completeness targets for critical variables (eg, birth year, sex) are >92%
	Patient level	Patient-level completeness is assessed by verification checks designed to identify and improve potentially incomplete data on the basis of clinical or data model expectations	Patients with a line of therapy change without a corresponding progression event are reviewed for complete capture of progression Patients who have received a PI3K inhibitor but do not have a PIK3CA test are reviewed for complete capture of biomarker test data
	Variable level	Variable-level completeness is assessed across a selected variable in a data set after curation, across sites, with thresholds set based on variable criticality and clinical or internal benchmarks	Completeness of smoking status, which is expected to be frequently captured in the EHR patient chart, has an expected completeness of >95%
	Field level	Field-level completeness is assessed by verification checks designed to identify and improve potentially incomplete data on the basis of clinical or data model expectations	Abstracted treatment start dates containing a year but missing month or day are re-reviewed for more complete data capture Required fields as per the data model, such as diagnosis, are prompted to be completed during data curation before submitting data with quality controls
Provenance	Data collection	Information about data sources, setting and time period of collection, and timing of extracts	Data elements can be traced to specific site, setting, extraction date, and source documentation Distributions of data source site (community cancer centers, academic medical centers), geographic areas, and patient populations are made available
	Processing	Information about the steps to curate and transform the source data	Abstractor username, policy version, and timestamp are logged for curation from unstructured data Version of the data standard used for mapping and mapping decisions are stored Data changes over time are logged with an audit trail
	Data and quality management	Documentation of processes for data and quality management	Data management plans are available and version controlled Training records for staff handling data are logged and retained Data verification checks are version controlled, with records of flagging and resolutions
Timeliness	Recency-based thresholds	Percent of patients with a value within a given window of time	Percentage of nondeceased patients with a medication administration in the 90 days before data cutoff
	Data refresh cadence	Frequency with which incremental documentation is curated within the data set	Structured data feeds are refreshed every 24 hours

Abbreviations: EHR, electronic health record; PI3K, phosphoinositide 3-kinases; PIK3CA, phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha.