NIH PRAGMATIC TRIALS COLLABORATORY
Rethinking Clinical Trials®

Assessing Fitness-for-Use of Clinical Data for PCTs

Background

The credibility and reproducibility of pragmatic clinical research depends on the investigator's demonstration that the data are of sufficient quality to support the research conclusions. This document highlights recommendations for assessing the fitness-for-use of data generated from routine patient care for use in pragmatic clinical trials (PCTs). For more information, read the full chapter in the Living Textbook: Assessing Fitness for Use of Real-World Data.

Before using an electronic health record (EHR) dataset for a given research project, one should determine whether it is fit-for-purpose by determining if the data are **relevant** and **reliable**. Relevance includes the availability of key data elements (exposures, outcomes, covariates) and sufficient number of representative patients for the study. Reliability includes data accuracy, completeness, provenance, and traceability (FDA 2021).

More specifically, a real-world data source is said to be **relevant** if:

- The data apply to the question at hand
    - For example, the data contain sufficient detail to capture the use or exposure of the product or device and/or the outcome of interest
- The data are amenable to sound clinical and statistical analysis
    - For example, the data can be used to answer the specified question using the proposed statistical plan
- The data and evidence the source provides are interpretable using informed clinical and statistical judgement.
    - For example, the use of a device or product in a real-world population is representative of what is captured in the data source, is generalizable to the relevant population under study, etc. (FDA 2018).

Data are considered **reliable** if:

- Data are captured in a standardized and rigorous manner
- Data are accurate and complete, data provenance is known, and data are traceable
- Efforts of data curation, transformation, accrual, etc. are known (i.e., process from transforming raw data to analytic dataset)

EHR data typically go through several phases when used to support a PCT—from source system, to clinical data repository to data warehouse to study-specific dataset. The quality or fitness of a dataset may be evaluated at various points along this process, with different processes for quality assurance or quality control (FDA 2021). Assessment of data quality is on ongoing process, and conformance, completeness, and plausibility should be assessed throughout the trial.

Data Quality Checks

Example data checks to evaluate conformance, completeness, and plausibility are provided in the table below.

Table 1. Categories of Data Quality Checks and Examples From Distributed Research Networks

Category	Subcategory	Description	Data Check Example
Conformance	Value	Determines whether the data conform to the formats of the data model used to store them	Sex values are F, M, or U; age is in specified range
	Relational	Determines whether the data agree with the constraints imposed by the database used to store them (e.g., primary or foreign key relationships)	All patient medical record fields are present in each table that requires them
	Calculation	Evaluates whether variables derived computationally yield valid results	Enrollment periods do not overlap; computed BMI is correct
Completeness		Examines whether expected values are present (single time point or longitudinally)	Gender is not null
Plausibility	Uniqueness	Determines whether multiple values exist when only one value is expected	Patient does not have multiple inpatient admissions to the same facility on the same day
	Atemporal	Measures whether data agree with expected values	Most of the records are not in the lowest or highest categories of age, height, weight, diastolic blood pressure, etc.
	Temporal	Examines whether variables change as expected over a specified time period	Events are not before date of birth or after date of death

For more details see: A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data and the FDA Guidance for Industry, Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products

Data Quality Assessment Recommendations for PCTs

1 – Key data quality dimensions

We recommend that conformance, completeness, and plausibility be formally assessed for data elements used in subject identification, outcome measures, and important covariates.

2 – Reporting data quality assessment with research results

Results of data quality assessments should be reported with research results. Data quality assessments are the only way to demonstrate that data quality is sufficient to support the research conclusions, and as such should be accessible to consumers of research.

Food and Drug Administration. 2018. Framework for FDA's Real World Evidence Program. https://www.fda.gov/media/120060/download. Accessed July 12, 2022.
Food and Drug Administration. 2021. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products. https://www.fda.gov/media/152503/download. Accessed July 12, 2022.

The NIH Pragmatic Trials Collaboratory is supported by the National Institutes of Health (NIH) through cooperative agreement U24AT009676 from the Office of Strategic Coordination within the Office of the NIH Director. It is also supported by the NIH through the NIH HEAL Initiative under award number U24AT010961. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or its HEAL Initiative. To learn more about the program, visit rethinkingclinicaltrials.org.