Determining Real-World Data's Fitness for Use and the Role of Reliability

September 26, 2019

Duke | MARGOLIS CENTER *for* Health Policy

Table of Contents

CHAPTERS:

Chapter 1: Unpacking Fitness-for-Use Concepts — The Role of Verification Checks to Assess RWD Reliability 6

Chapter 2: Special Considerations for Verification Checks to Assess EHR Data Reliability 17

Chapter 3: Special Considerations for Verification Checks to Assess PGHD Reliability 21

APPENDICES:

Appendix A. Workshop Participants 29

Appendix B. Glossary.....31

Appendix C. RWD Curation 33

Appendix D. Approaches to RWD Curation Processes 36

Appendix E. Verification Checks to Assess Data Reliability 37

Appendix F. Data Linkage.....38

Appendix G. Electronic Health Records and Meaningful Use 39

Appendix H. EHR Breakout Session Tool 40

Appendix I. Relational Databases 44

Appendix J. PGHD Breakout Session Tool 45

FIGURES:

Figure 1. Fit-for-Use RWD Framework 7

Figure 2. Examples of Quality Check Principles 9

Figure 3. How Verification Checks for Completeness, Conformance, and Plausibility Contribute to Assessing Reliability to Support Fit-for-Use RWD.....10

Figure 4. The Four Types of PGHD 23

Figure 1A. Single-Stage Data Curation Process.....34

Figure 1B. Two-Stage Data Curation Process.....35

About the Duke-Margolis Center for Health Policy

The Robert J. Margolis, MD, Center for Health Policy at Duke University is directed by Mark McClellan, MD, PhD, and brings together expertise from the Washington, DC, policy community, Duke University, and Duke Health to address the most pressing issues in health policy.

The mission of Duke-Margolis is to improve health and the value of health care through practical, innovative, and evidence-based policy solutions.

Duke-Margolis catalyzes Duke University's leading capabilities, including interdisciplinary academic research and capacity for education and engagement, to inform policy making and implementation for better health and health care.

For more information, visit healthpolicy.duke.edu.

Acknowledgments

The Duke-Margolis Center would like to thank several individuals for their contributions to this white paper. This paper's development was guided by the collaboration of the working group members listed on page 2. This paper would not have been possible without their willingness to participate in open discussion, thoughtful comments, and extensive subject-matter expertise. We are grateful for the support of the working group and the broader Real-World Evidence Collaborative Advisory Group who provided their expertise throughout development of this white paper. We would also like to thank those who participated in the Duke-Margolis private workshop, "Principles for Developing Quality Checks to Assess Fit-for-Use RWD," held in May 2019 (a list of attendees can be found in Appendix A). In addition, we are grateful to Patricia Green and Sarah Supsiri from the Center for their communications and events guidance. Finally, we acknowledge Kerry Stenke from the Duke Clinical Research Institute for her support in developing graphics for this white paper, and her colleague, Damon Seils, for editorial assistance. Any opinions expressed in this paper are solely those of the authors and do not represent the views or policies of any other organizations external to Duke-Margolis.

Funding for this work is made possible through the generosity of the Margolis Family Foundation, which provides core resources for the Center, as well as a combination of financial and in-kind contributions from Real-World Evidence Collaborative members, including AbbVie; Eli Lilly and Company; Genentech, a member of the Roche Group; GlaxoSmithKline; Johnson & Johnson; Merck & Co.; Novartis; Teva; and UCB. For more information on the Real-World Evidence Collaborative, visit https://healthpolicy.duke.edu/real-world-evidence-collaborative.

Disclosures

Mark B. McClellan, MD, PhD, is an independent board member for Johnson & Johnson, Cigna, and Alignment Health Care; co-chairs the Accountable Care Learning Collaborative and the Guiding Committee for the Health Care Payment Learning and Action Network; and receives fees for serving as an advisor for Cota and MITRE.

Gregory Daniel, PhD, MPH, receives consulting fees from Genentech, Inc., the Reagan-Udall Foundation, and AbbVie Inc.

Rachel E. Sherman, MD, MPH, is an independent board member for Aptinynx and has served as an advisor to Medidata.

WHITE PAPER

Determining Real-World Data's Fitness for Use and the Role of Reliability

DUKE-MARGOLIS AUTHORS

Nirosha Mahendraratnam

Christina Silcox

Kerra Mercon

Adam Kroetsch

Morgan Romine

Nicholas Harrison

Adam Aten

Rachel Sherman

Gregory Daniel*

Mark McClellan

* Duke-Margolis affiliation ended on August 2, 2019.

WORKING GROUP

Dongmu Zhang
AbbVie

Brad Hammill
Duke University

Keith Marsolo
Duke University

Michael Pencina
Duke University

Lisa Wruck
Duke University

Brande Ellis Yaist
Eli Lilly and Company

Stacy Holdsworth
Eli Lilly and Company

Luca Foschini
Evidation Health

Ernesto Ramirez
Evidation Health

Jason Simeone
Evidera

Sailu Challapalli
Flatiron Health

Laura Koontz
Flatiron Health

Grazyna Lieberman
Genentech

Anne-Marie Meyer
F. Hoffmann-La Roche

Shannon Ferrante
GlaxoSmithKline

Sulabha Ramachandran
GlaxoSmithKline

Kevin Haynes
HealthCore

Marni Hall
IQVIA

Christina Mack
IQVIA

Marc Berger
International Society for Pharmacoeconomics and Outcomes Research

Margaret (Clair) Blacketer
Janssen Research and Development

Vinay Mehta
Merck & Co.

Ellen Snyder
Merck & Co.

Barbara Bierer
The Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard

Kristijan Kahler
Novartis

Aylin Altan
OptumLabs

Paul Bleicher
OptumLabs

David Blaser
PatientsLikeMe

Sally Okun
PatientsLikeMe

Marko Mychaskiw
Teva Pharmaceuticals

Tim Sampson
UCB

ADVISORY GROUP

Ryan Kilpatrick
AbbVie

Carlos Garner
Eli Lilly and Company

Bray Patrick-Lake
Evidation Health

Debra Schaumberg
Evidera

Nicole Mahoney
Flatiron Health

Jacqueline Law
Genentech

John Graham
GlaxoSmithKline

Richard Platt
Harvard Medical School

Marcus Wilson
HealthCore

Marc Berger
International Society for Pharmacoeconomics and Outcomes Research

Richard Willke
International Society for Pharmacoeconomics and Outcomes Research

Nancy Dreyer
IQVIA

Joanne Waldstreicher
Johnson & Johnson

Solomon Iyasu
Merck & Co.

Barbara Bierer
The Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard

Eleanor Perfetto
National Health Council

Amy Rudolph
Novartis

Paul Bleicher
OptumLabs

Sally Okun
PatientsLikeMe

David Thompson
Syneos Health

Riad Dirani
Teva Pharmaceuticals

David Miller
UCB

Stephanie Reisinger
Veradigm Health

Observers

Jacqueline Corrigan-Curay
US Food and Drug Administration

Amanda Wagner-Gee
National Academies of Sciences, Engineering, and Medicine

EXECUTIVE SUMMARY

Evaluating whether real-world data (RWD) are fit for use is critical for generating real-world evidence (RWE) to inform regulatory decision making on the effectiveness of medical products. **This paper describes a framework for how researchers and reviewers can systematically evaluate whether RWD are fit for use by using verification checks to assess reliability.** In Chapter 1, we identify key concepts that should be evaluated in assessments of reliability, including completeness, conformance, and plausibility. Chapters 2 and 3 focus on considerations for applying the framework to electronic health record (EHR) data and person-generated health data (PGHD), although the framework may also be applied to other sources of data, such as administrative claims and billing data (claims data) and patient registries.[*] This paper aims to inform the global health care research community on data reliability, serving as a resource for sponsors as they design studies using RWD sources, for regulators as they develop policy, and for researchers as they develop best practices for study methods. Of note, the US Food and Drug Administration (FDA) has identified issuing guidance on assessments of the reliability and relevance of RWD in generating RWE on drug product effectiveness as a program item in its 2018 RWE Framework.[1] This paper does not reflect the views, guidance, or recommendations of FDA and should not be construed as such.

Demonstrating RWD reliability is critical to establishing "whether the data adequately represent the underlying medical concepts they are intended to represent."[1] One way to systematically assess RWD reliability is to use a standardized set of verification checks to evaluate whether a data element or variable matches expectations with respect to metadata constraints and system assumptions within the dataset.[2] However, it may not be feasible to identify a standardized set of checks to assess **all** aspects of reliability due to heterogeneity within and between RWD. Instead, **a minimum set** of standardized verification checks used to assess **some** aspects of reliability across all data sources could be identified and adopted as a first step. Because data curation is a dynamic process, this minimum set of verification checks should be assessed continuously based on initial review and on findings during analysis. The process used to assess and address data reliability should be prespecified in research protocols and/or statistical analysis plans.

The minimum set of verification checks to assess data reliability could be used by researchers to differentiate data sources that have the potential to be fit for regulatory use from those that do not, and to serve as a starting point for reviewers to evaluate real-world datasets submitted to FDA as part of evidence packages. However, additional verification checks specific to the research question and the

[*] While claims and registry data are also important sources of RWD, they were not included in this paper because there is already a more advanced understanding of how to characterize their quality.[1] Lessons learned from verification checks used to evaluate claims data reliability informed the development of this paper, and identifying a minimum set of claims data-specific checks is a potential topic for future work.

data will generally be required. "Failure" of a reliability check does not necessarily render the data unusable, so long as the underlying problem can be addressed or the impact on the research question can be explained.

Using verification checks to assess data reliability is a critical step in determining whether RWD are fit-for-use, but this step alone is not sufficient. Fitness for use also depends on data relevancy and other aspects of reliability, including validation.[2,3] In addition, whether the RWE can adequately answer the regulatory research question depends on the methods used as well as the regulatory and clinical contexts.[4]

How This Paper Was Developed

This paper is informed by a literature review, a full-day private workshop on "Principles for Developing Quality Checks to Assess Fit-for-use RWD" (May 28, 2019), and the expert opinion of the Duke-Margolis RWE Collaborative RWD Quality Working Group. During the workshop, stakeholder experts representing sponsors, payers, research groups, data vendors, providers, and patient networks provided feedback on a list of reliability checks. A modified version of the PCORnet Data Quality Checks (Version 6) was used to guide discussions because it is publicly available, it contains credible content developed by data quality experts, and it is succinct. Recommendations as to the specific verification checks that should be used to assess data reliability are not made in this paper. Instead, key concepts and considerations to identify a minimum baseline set of checks along with example checks are highlighted. This work builds on prior Duke-Margolis work, including the white papers *Characterizing RWD Quality and Relevancy for Regulatory Purposes* (2018) and *A Framework for Regulatory Use of Real-World Evidence* (2017).[3,4*]

[*] Because both of these papers were published prior to the release of the 2018 FDA RWE Framework, some terms and concepts have been modified for alignment (e.g., the concept of data reliability in this paper is termed *data quality* in the previous papers).

Chapter 1: Unpacking Fitness-for-Use Concepts — The Role of Verification Checks to Assess RWD Reliability

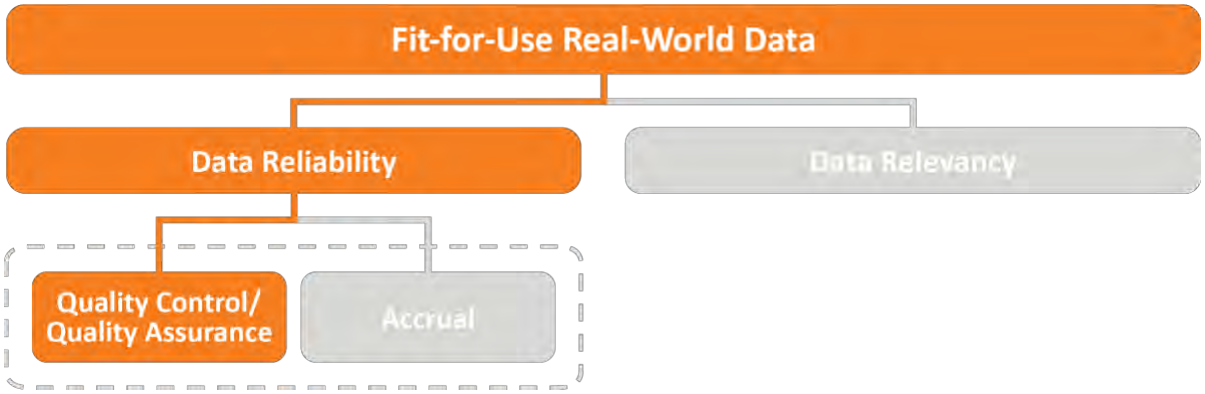
Background

Stakeholders are eager to leverage real-world data (RWD), or "data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources," into drug development and regulatory decision making.[1] In particular, there is interest in analyzing RWD to generate real-world evidence (RWE) about the potential benefits or risks of medical products.[1] Compared with evidence generated through traditional clinical trials, RWE may be more inclusive of broader patient populations, including subpopulations, and may better reflect routine clinical and self-care.[1] RWE may also offer evidence that is generated more efficiently. Fulfilling a requirement set forth in the 21st Century Cures Act of 2016, the US Food and Drug Administration (FDA) released the framework on its Real-World Evidence Program in December 2018. The framework focuses on considerations for using RWD and RWE in regulatory decision making and includes a three-pronged approach that examines whether "(1) RWD are fit for use; (2) studies that use RWD can provide adequate scientific evidence to answer regulatory questions; and (3) study conduct meets regulatory requirements."[1]

Determining the appropriate use of RWD requires modernization of clinical trial approaches to incorporate the increasing digitalization of health data as well as new and novel technologies. In particular, evaluating whether RWD are fit for regulatory use is complex because (1) it depends on the research question; (2) the data were not generated for the purpose of conducting research (e.g., secondary data use); (3) there can be substantial heterogeneity within and between RWD sources in terms of how data are captured and the types of information collected; and (4) there is a lack of universal standards in data curation and measures of fitness. RWD evaluation is also complicated because a single source of RWD is often unlikely to be sufficiently fit for use on its own and may need to be linked to other sources. Additional complications may arise when multiple RWD sources are used to address the same research question and when RWD sources can only partially address the research questions and supplementary primary data are needed.

Fit-for-use RWD is a multifaceted concept.[1] It implies that the data meet a standard establishing that the data and the resulting analysis of it can be successfully used to inform regulatory decision making (e.g., labeling changes) (Figure 1). Fit-for-use RWD are data that are reliable and relevant to the regulatory research question.[1]

Figure 1. Fit-for-Use RWD Framework[*]



Data reliability "considers whether the data adequately represent the underlying medical concepts they are intended to represent."[1] Data relevance is an "assessment of whether the data can adequately address the regulatory question, in part or whole" (e.g., not using a lupus-specific database to answer a pancreatic cancer question, or using a database with only all-cause mortality to study disease-specific mortality).[1,5] Data reliability addresses whether the data are trustworthy and credible. RWD reliability is demonstrated, in part, through data accrual and data quality control/quality assurance (Figure 1). Data accrual refers to how data are collected, and quality control/quality assurance focuses on whether the "people and processes in place during data collection and analysis provide adequate assurance that errors are minimized and that the data quality and integrity are sufficient."[1,5] (These terms and others are defined in the glossary in Appendix B.)

Data Accrual

"To ensure the reliability of RWD, the RWD source should have an operational manual or other documentation that pre-specifies the data elements to be collected, data element definitions (i.e., data dictionary to provide a common definitional framework), methods for data aggregation and documentation (e.g., common case report form, abstraction from verifiable sources), and the relevant time windows for data element collection (i.e., common temporal framework). Some RWD sources such as EHRs or claims data may not fulfill all of these characteristics, but still be sufficiently reliable for regulatory decision making."[5]

The terms **quality control** and **quality assurance** are often used in tandem, but they represent distinct concepts. Quality control consists of the steps taken during data curation to ensure that data meet prespecified standards and that they are reproducible.[6] For example, quality control might include the steps taken to convert nonnumerical age data into numerical data. Quality control practices should be specified, documented, and justified to ensure a rigorous process. Quality assurance consists of

[*] Figure developed based on FDA's RWE Framework (2018).[1] This paper focuses on the reliability aspect of fit-for-use RWD.

proactive and retrospective steps undertaken to evaluate whether prespecified requirements are fulfilled.[6] For example, quality assurance might include running descriptive statistics on the metadata of an age variable and providing the percentage of date variable data that were in numerical format, both at the raw data stage and after quality control practices were applied.

Data Assurance[*]—Quality Control

"Data quality control is essential for providing confidence in the reliability of RWD and RWE sources… However, certain sources of RWD, such as some administrative and health care claims databases or EHRs, may not have established data quality control processes and may not be capable of fully implementing or following the above recommendations. When considering a source of RWD for regulatory purposes, it is important to consider any methods and systems used to help ensure sufficient data quality. Potential RWD sources should be evaluated in accordance with the data [quality assurance] plan and procedures developed for the data source itself."[5]

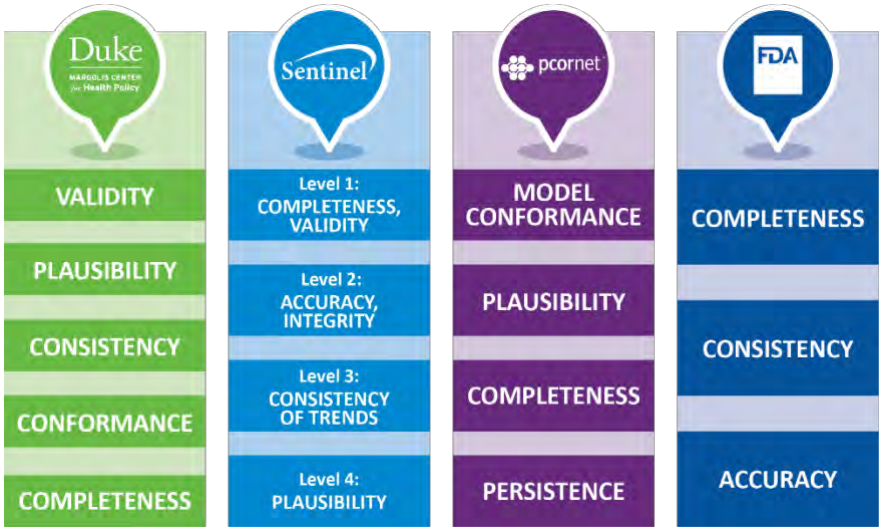
Deriving a Baseline Set of Verification Checks to Assess Data Reliability

Transparency and assessment of data curation practices are necessary to allow regulators to properly evaluate whether RWD submitted for regulatory review are of sufficient reliability (see Appendix C for more information on RWD curation). Many approaches to assessing RWD curation exist, including identifying best practices (e.g., quality control), evaluating documentation of quality control practices, and/or evaluating quality assurance practices. For each of these data curation assessment practices, we evaluated the feasibility of implementation among data vendors and sponsors and the resulting interpretability and review burdens on FDA (see Appendix D for more information). **However, given the difficulty of identifying a common set of data curation practices that could be evaluated systematically and the resource-intensiveness of reviewing documentation, the ability of the regulators to determine a dataset's fitness for use could be facilitated by identifying key considerations for a set of quality assurance practices, specifically verification checks**. **In this case, the checks would provide information on the real-world dataset using statistical measures or standardized summary documents to show whether the data transformation suitably addressed data reliability characteristics.** Because data curation is a dynamic and iterative process, checks should be assessed over time and throughout the data curation process (e.g., initial data ingestion, cleaning, transformation).

[*] The term "data assurance" is used in FDA's *Guidance to Industry on the Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices*, whereas the term "quality assurance" is used in the *Framework for FDA's Real-World Evidence Program*. Our interpretation of the uses of these terms in the context of their respective papers is that they represent similar concepts. We use the term "quality assurance" in this paper for consistency.

To understand the landscape of checks used to assess reliability (often referred to as "data quality" in the literature), multiple frameworks and checks used by data aggregators* were reviewed and compared (Figure 2).[2,3,5-9] While similar checks were used across these data sources, the lack of standardization across data reliability principles complicated the identification of a core set. As a result, a study published by Kahn et al. that proposed a conceptual model centered on harmonizing EHR data quality (reliability) principles that is also applicable to other RWD sources was leveraged.[2] This study was informed by, but not limited to, standard operating procedures for data quality, data quality publications, and expert opinion from distributed research networks, such as FDA's Sentinel Initiative and the Observational Health Data Sciences and Informatics (OHDSI) program, as well as large dataset aggregators. The authors proposed a broad framework identifying key data quality (reliability) principles to help end users assess whether a dataset could meet its intended use. In the Kahn et al. harmonization framework, data quality (data reliability) can be assessed by verification or validation of conformance, completeness, and plausibility (Figure 3).[2†] **Verification checks evaluate whether the data element or variable matches expectations with respect to metadata constraints and system assumptions within that dataset.**[2] While data verification is a common concept in engineering (including medical devices) and computer science, its application in epidemiology is newer but is especially relevant to evaluating RWD reliability.[10] Conversely, validation checks focus on comparison of the data element or variable to another data source (e.g., an external gold standard).[2]

Figure 2. Examples of Quality Check Principles



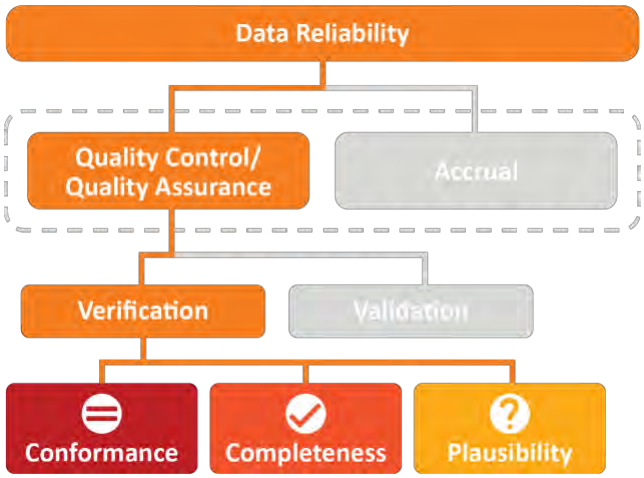
Subsequently, it was decided to prioritize the understanding of verification checks to assess reliability, rather than validation checks as the first step to standardizing the assessment of RWD fitness for use for two reasons (Figure 3). First, verification checks to assess reliability are more prevalent. Callahan et al. mapped checks used by six data sharing networks, including OHDSI and Sentinel, to the Kahn et al.

* The literature review included, but was not limited to Sentinel and PCORnet resources, FDA's Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Datasets, and Standards for Data Management and Analytic Processes in the Office of Surveillance and Epidemiology (FDA MAPP 6700.2).

† While *quality control/quality assurance* and *accrual* are described discretely within FDA's Fit-for-Use Framework, as illustrated in Figure 1, the concepts are not mutually exclusive. As such, Figure 3 illustrates *quality control/assurance* and *accrual* as related concepts (dotted line).

framework and found 99 percent of over 11,000 checks aligned with the framework.[2,11] While validation checks also will be needed to determine fitness for use, only 15 of those 11,000 checks were for validation.[2,11] Second, it is often challenging or not possible to identify a comparable gold standard to test validation.

Figure 3. How Verification Checks for Completeness, Conformance, and Plausibility Contribute to Assessing Reliability to Support Fit-for-Use RWD*



Fit-for-Use Data: Using Verification Checks to Assess Reliability

Following the framework proposed by Kahn et al., this section discusses considerations for identifying a minimum set of verification checks to assess data reliability, specifically conformance, completeness, and plausibility.[2] Assessing data reliability through the minimum set of verification checks is a critical step in determining whether data are fit for use, but this step alone is not sufficient. First, supplemental verification checks specific to the research question as well as validation checks will be required to fully assess RWD reliability. Second, what level of RWD reliability is considered acceptable will be on a case-by-case basis.[2-4,12] For example, one regulatory research question may tolerate a higher threshold for missingness of a particular variable than another regulatory question. Third, RWD fitness for use also depends on data relevancy. (In addition, whether the RWE can adequately answer the regulatory research question depends on the methods used as well as the regulatory and clinical contexts.[4])

For data source–specific considerations for EHRs and PGHD, see Chapter 2, "Special Considerations for Verification Checks to Assess EHR Data Reliability," and Chapter 3, "Special Considerations for Verification Checks to Assess PGHD Reliability."

Conformance Verification Checks

Verification checks for data conformance assess the structure of the data and how compliant the data are with internal relational, formatting, or computational definitions or standards.[2] In other words, conformance checks indicate compliance with a prespecified standard. Conformance may be implemented through constraints, which define the properties with which data must comply.[13] For structured data, conformance checks can be derived from (1) database-level constraints dictated by the physical architecture of the database; (2) data model standards such as variable or data element specifications on the type, range, or formatting (often found in data dictionaries or codebooks);

* Figure adapted from FDA's RWE Framework (2018) and Kahn et al.'s Data Quality Harmonization Framework (2016).[1,2]

or (3) alignment between manually computed and programmed variables (e.g., calculated body mass index [BMI] using height and weight is the same as the BMI variable).[2]

Data in static or dynamic databases can be made "research-ready" by standardizing and defining the data through a codebook, data model, or predefined abstraction process. However, each database owner will make their own decisions about standardization, even when working from the same data source. For example, EHR data aggregators who have direct control or access to the underlying source systems may rely on metadata to translate between study concepts and the underlying data models. Distributed data networks, on the other hand, may ask data holders to harmonize their source data into a common data model to allow the use of a common query or analysis. In both cases, differences between EHR platforms increase the complexity of assessing data reliability. A few open-source common data models include the Sentinel Common Data Model, the Observational Medical Outcomes Partnership (OMOP) Common Data Model, and the PCORnet Common Data Model, among others. Common data models are more available in claims and EHRs than PGHD.

Conformance checks can also apply to unstructured data. Unstructured data (e.g., free text) can provide information from which structured concepts can be extracted or derived and then mapped to structured elements. The results of such extraction and mapping can be characterized by conformance to an expectation or standard. For example, users may want to understand the degree to which unstructured data from free text can be interpreted by either human or machine consumers. A conformance check might then characterize interpretability by comparing the output of mapping to standards regarding the percentage of sentences that cannot be atomically parsed by syntax, contain unknown or misspelled words, are very long or short, or contain special characters, URLs, or other nonstandard characters.[14] Ultimately, the provenance of the data and the processes of generating structured concepts or data that are derived from unstructured data should be well documented, such as within study protocols and statistical analysis plans.

Regardless of whether the raw RWD are stored as structured or unstructured data, maintaining the integrity of the raw data and ensuring the availability of documentation that identifies the origin of the data and the history of all data transformations (e.g., data provenance and data lineage) are important as data are mapped and transformed into a fit-for-use dataset.[15] Adequate data documentation preserves end users' ability to check their mapping and transformations, both of which should be prespecified and justified. This documentation may be especially useful if the transformations result in output that is out of context for some users, even if it is appropriate for other users' purposes. To preserve raw data in the absence of source documents or a saved snapshot of the raw data, it may be useful to pursue data characterization—summarizing its general features—where otherwise attempting to force data into a model might result in data exclusion. In addition, raw data should never be directly manipulated; instead, transformations should be traceable and documented.

Completeness Verification Checks

Verification checks for data completeness assess "the frequencies of data attributes present in a dataset without reference to data values."[2] In other words, a check for completeness verifies the presence of a value rather than the value itself. Completeness checks measure how a data environment aligns with expectations (e.g., the death rate in EHR data is expected to be low). Accordingly, completeness checks commonly depend on pragmatic thresholds based on context that indicate where datasets may have reliability faults. Thresholds should be evidence-based, justified, and defined, ideally before the analysis, but can change over time. For example, the presence of emergency department events in an EHR can be expected to differ between source data from a hospital and source data from a primary care provider group. While a completeness check indicating an unexpectedly low rate of emergency department visits in a hospital data source may be cause for concern, the same check applied to a dataset from a primary care group may not raise concern. As a practical example, PCORnet includes completeness checks that raise awareness when less than a certain percentage of patients with encounters have diagnosis records (PCORnet check 3.04) and when the average number of diagnosis records with known diagnosis types is below the specified threshold (PCORnet check 3.01).[8] (All checks referenced in this paper can be found in Appendix E.)

RWD are subject to different types of missingness, which impact the ability to make causal inference.* It is necessary to characterize what is missing, the rate of missingness, and why it is missing (e.g., incomplete mapping versus variable not captured) to understand their type. Understanding their type can inform whether biases exist and can be adjusted for in the study design and analysis.[17] For instance, differentiating between missingness due to an event that occurred that was not captured in the RWD source and missingness due to an event not occurring that is expected based on standard of care. An example of the first type of missingness is heart rate data not captured from a wrist-worn heart rate tracker when the patient has removed the device for charging or during bathing. The device is expected to record 1440 minutes of beat-per-minute values (i.e., 24 hours multiplied by 60 minutes). However, because patients will regularly remove the device, analysis of the data may only require 600 minutes of data per day. Completeness checks in this situation should consider the amount of data required for the analysis, which is typically determined a priori. In an example of the second type of missingness, recording vitals during ambulatory encounters may be specified in guidelines, but height may not be recorded twice for an adult with two ambulatory encounters in the same month.

When completeness checks flag reliability concerns, investigators may be unable to use the data, need to seek additional or alternate data, need to impute data, or need to modify the statistical analysis. In

* Missingness can be systematic (e.g., all scanned documents are excluded from an EHR) or unsystematic (only scanned documents from one provider are excluded from an EHR due a random software issue) and can influence study design considerations. The different types of missingness include missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR).[16] If missing data are unrelated to the data for all other variables or to non-missing data of the same variable, it is MCAR. However, MAR is the more likely scenario, where the values of missing data depend on some other variable(s), but do not depend on non-missing data of the same variable. Otherwise, if the probability of missing data depends on the variable itself, it is said to be NMAR.

combination with context and institutional knowledge, completeness checks can indicate where data curation, study design, or analysis may need modification to address missing data.

Plausibility Verification Checks

Verification checks for plausibility assess the "believability or truthfulness of data values."[2] Plausibility can be evaluated by examining (1) the uniqueness of values; (2) the range and distribution of values within a variable, and whether two or more variables have an expected context-dependent relationship; or (3) whether time-related and time-varying variables change as expected. Verification checks for plausibility frequently depend on contextual knowledge and expectations regarding time-dependent relationships. Uniqueness as assessed by conformance checks occurs in the context of keying (e.g., each medical record number is assigned to one patient only). In contrast, uniqueness in plausibility checks is evaluated in the context of the value itself (e.g., each patient has only one medical record number).[2]

Plausibility checks evaluate data values against a standard that is considered believable. The range and distribution of values can indicate where plausibility issues may exist. For example, height observations are expected to have a normal distribution within certain ranges for specified age groups. The Measurement to Understand Reclassification of Disease of Cabarrus/Kannapolis (MURDOCK) Study Community Registry performs a check to flag height observations outside the range of 36 to 84 inches.[11,18] Other such checks may assess the distributions of weight, age, blood pressures, diagnostic values, and other common continuous variables. PCORnet checks whether more than 10 percent of records fall into the lowest or highest categories of age, height, weight, etc. (PCORnet check 2.02).[8] It may also be useful to leverage known relationships between multiple variables to assess plausibility. For example, plausible value ranges may not be consistent across pediatric and adult populations. In another example, prostate conditions are not expected to occur in women; likewise, female infertility is not expected to occur in men. In OHDSI, there are hundreds of sex-specific concepts in the Systematized Nomenclature of Medicine-Clinical Terms (SNOWMED CT) and the International Classification of Diseases, Tenth Revision (ICD-10) within the gender domain that can be checked against the recorded biological sex.[19] Context-dependent relationships are numerous and present various opportunities to evaluate plausibility.

Times and dates are fundamental to several checks for plausibility. Checking for implausible dates may be immediately useful, such as surveying when more than 5 percent of records have future dates (PCORnet check DC 2.01).[8] An important subset of context-dependent relationships are time-dependent relationships between variables. Expectations about these relationships may include that they vary or persist over time or that events proceed in a sequential order.[2] Each of these expectations can underlie a verification check for plausibility. For example, observational records might be expected to increase over time if a dataset represents a growing population or has incorporated new data streams. Kaiser Permanente's Center for Effectiveness and Safety Research (CESR) counts the number of observations by year across each year of data.[11,20] An example check for sequential order might check whether interactions were recorded before a sensor was turned on. Investigating sequence may raise flags even

where plausibility exists, as when autopsies are recorded after death or when future appointments are included in datasets. As with completeness checks, context is critical in determining whether verification checks uncover truly implausible data.

Although implausible data may represent only a fraction of the total data, they may be partitioned into context-explainable noise and unexplainable noise. For example, while there may be many instances of "noise" or unexpected variation in a variable, the reason for that noise may be singular. Data that are explainable can then be organized into useful RWD.

The Need for Transparency of Data Provenance, Accrual, and Curation to Assess Reliability

Researchers receive varying levels of curated data (i.e., raw data versus research-ready data), and the ability to apply verification checks to assess reliability depends on understanding the data curation process (Appendix C). Knowledge of RWD provenance or the origin of data is critical, so that the researcher or curator can return to the source data if reliability issues arise. It is also beneficial to know how data is transformed (i.e., data lineage) to identify what step(s) in the curation process may have created or exacerbated reliability problems.

The level of access to the entirety of the raw data by the researcher may vary and subsequently influence the ability to create and evaluate a fit for use dataset. This visibility depends on whether the data owner is curating the data (full visibility) or a third party is aggregating the source data under restrictions the data owner places on the aggregator (e.g., data use agreements). A researcher working with raw data is able to manage and document all curation steps and their rationale to subsequently report on verification checks. In contrast, a researcher using a research-ready database must rely on the third-party data curator to provide this information. (The curator could provide the researcher with documentation so that the researcher can report on the checks, or report on the checks themselves.) Accessing the level of information needed at the regulatory level from commercial third-party aggregators to evaluate fitness for use may be difficult. For example, many curation practices, while rigorous and scientific, are proprietary. A process that facilitates shared roles and responsibilities between researchers and data aggregators is necessary. Such a process needs to balance transparency with trade secrecy to maintain commercial interests.

RWD sources constantly accrue data as new patients and information are added through data refreshes. Data persistence describes or measures the data preserved and added between data refreshes. Evaluations of persistence determine whether increases or decreases in records, referred to as the "growth rate," are expected within a particular time frame. Data persistence may be affected for reasons other than the reliability of the data; for example, data licensing agreements may call for research sites to be dropped along with their associated data, affecting the persistence but not necessarily the reliability of the remaining data.

When aggregating data sources or tables within a data source, evaluating the reliability after data linkage is vital, as incorrect linkage can lead to missing data elements or duplicated records. (For more information on data linkage, see Appendix F.[21]) For example, linking data by using indirect identification of a unique identifier, such as a proxy patient-matching algorithm based on sociodemographic factors, may result in false positives or false negatives. Poor data linkage can cause selection bias and confounding.[22]

What Are the Next Steps?

Verification checks for conformance, completeness, and plausibility are used in assessing RWD reliability and determining whether the data are fit for use for regulatory decision making. **Identification of a minimum set of verification checks that assess reliability** can help researchers and reviewers systematically evaluate RWD fitness for use.

Verification checks help characterize data and provide decision makers (e.g., regulators) and researchers with a better understanding of the data's potential limitations. Characterizing data rather than removing data with reliability issues allows researchers to determine whether the dataset is adequate for investigating a specific research question. These verification checks can also identify and differentiate between process failures (which are fatal to research) and content failures (which may be resolved depending on the requirements of the research question).

Oftentimes, numerous stakeholders are involved in collecting data and creating research-ready datasets derived from EHRs. From hospitals and care providers to third-party data aggregators and curators, the responsibility to maintain and communicate data reliability is shared, because the process of curating data affects its fitness for use. For this reason, the transparency of curation practices is critical. Data vendors that provide RWD for regulatory decision making should consider supporting transparent standard operating procedures (SOPs). Maintaining confidentiality where appropriate and communicating data reliability and curation practices is a necessary and shared responsibility.

A lack of best practices (or even standard practices) exists for communicating data fitness-for-use information in a manner that can be interpreted efficiently by FDA reviewers. Verification checks consist of a mix of quantitative outputs (e.g., descriptive statistics, graphs) and qualitative outputs (e.g., text descriptions of what is seen in the data). In addition, verification checks are assessed multiple times throughout the process of transforming raw RWD to a fit-for-use dataset. Providing all of this information multiple times throughout the data curation process, while necessary to understand, can be burdensome to review. At the workshop, participants suggested that another level of synthesis (i.e., a top-level summary of data reliability) presented through colors or Harvey balls, could be used to give reviewers an overall idea of data reliability. While such a summary may offer a quick assessment of data reliability, it may unintentionally bias the reviewer because of its lack of detail. More work in this area is needed, including identifying best practices for conveying the most relevant aspects of data reliability.

When quality assurance, quality control, and data accrual processes are consistent, there may be opportunity to precertify these processes to indicate when real-world datasets are research-ready for regulatory use. Such a certification would signal that appropriate SOPs and quality controls were used, limiting the number of times verification checks to assess reliability must be repeated on the same dataset between refreshes. While it may be possible to precertify data at a certain point in time or for a certain time period when they are not changed, precertification will largely depend on a dataset's intended use. Precertification could provide a level of confidence that suggests a dataset is of high reliability, but it does not eliminate the need for evidence to show that a dataset is fit for use to investigate a specific research question for regulatory decision making.

Defining a minimum set of verification checks to assess reliability is the first step toward assessing whether RWD are fit for use for regulatory decision making. Such checks provide a foundation for researchers and FDA reviewers to consider when evaluating RWD as part of evidence submissions. However, new research methods, development of advanced analytics and data curation techniques, and evolution of data continuously change the landscape. Thus, identifying a consistent mechanism to assess the reliability of RWD and determine whether they are fit for use requires an iterative approach, ready to incorporate new insights as they arise.

Future topics to be explored to support RWD fitness for use include:

- identification of data curation best practices by data source to fulfill verification checks to assess reliability;
- fitness-for-use validation checks to assess reliability; and
- fitness-for-use relevancy checks.

Chapter 2: Special Considerations for Verification Checks to Assess EHR Data Reliability

Background

EHRs have evolved to capture and present detailed data regarding the events and interactions that occur as part of a patient's health care experience and provide opportunity to leverage RWD to generate RWE to inform regulatory decision making. EHRs contain structured and unstructured data fields that include patient demographic and health information from clinical encounters, including diagnoses, symptoms, treatments, test results, prescriptions, patient experience data, and clinical narratives.[*] In addition to these native data elements, EHRs include peripheral documents such as imaging data, pathology reports, and patient history documents. The 2009 Health Information Technology for Economic and Clinical Health (HITECH) Act's meaningful use standards advanced EHR implementation in the United States.[23†] Today, with approximately 70 percent of US physicians adopting them, EHRs offer a rich source of patient data, including expanding and increasingly granular data.[24] EHRs were developed to streamline workflows and support electronic billing in hospitals and medical offices and to increase efficiency in information retrieval, reveal gaps in care, and advance clinical decision support tools.

As EHR platforms advance, the manner in which data are captured, stored, transformed, and shared has created opportunities for secondary-use EHR data to answer regulatory-grade research questions on product effectiveness. To ensure their fitness for use for regulatory decision making, the reliability of EHR data must be assessed and appropriately characterized as a first step toward fit-for-use databases. In this chapter, specific context for EHRs is provided based on discussions at the workshop. At the workshop, the EHR breakout session used the PCORnet Data Quality Checks as a baseline list of verification checks to assess reliability to spur discussion, as these checks are fairly advanced for EHR data (Appendix H). Due to heterogeneity in terms, platforms, and networks, an applicable minimum set of checks was not extracted, but rather principles for how to assess the reliability of a dataset were discussed.

Context for Assessing Reliability in EHRs

Due to workflow variety and the evolution of EHR platforms, EHR data attributes, such as format, terminology, and storage/organization, vary between health systems and can even evolve over time within a health system. As such, these data attributes must be standardized for analysis and evaluation. For example, not all formats of unstructured data (e.g., free text) are universally machine-readable and

[*] The Public Health Service Act definition of electronic health record can be found in Appendix G.
[†] The HITECH Act was enacted as part of the American Recovery and Reinvestment Act of 2009. The "meaningful use" language from the HITECH Act can be found in Appendix G.

may require human interpretation, thus hindering data extraction. Lack of standardization in terminology complicates data extraction for a particular clinical concept (e.g., all diagnostic lab results, values, and images associated with a cancer diagnosis) and requires harmonization by data aggregators. Another challenge is that peripheral documents are stored as attachments that may not be adequately accessible by data aggregators, let alone in a format from which data can be abstracted.

Conformance Verification Checks for EHRs

Verification checks for data conformance assess the structure of the data and how compliant the data are with internal relational, formatting, or computational definitions and standards.[2] EHR data are often stored and/or analyzed using relational databases, which highlight specific considerations for conformance verification checks. Relational databases define and enforce prespecified data attributes and organization, preventing duplication, helping with data management through the automatic propagation of modified data elements, and limiting the opportunity for conformance-related issues. Database-level constraints underpin the physical architecture of relational databases (see Appendix I for more on relational databases). To assure this architecture, a subset of conformance checks can be used to characterize whether column- or table-level constraints have been properly applied. Typical constraints include the presence of primary keys and foreign keys, as well as unique and not null constraints.* In addition, depending on the particular EHR, some fields may be optional and thus the number of primary keys may be small or consist only of a unique record identifier. PCORnet incorporates a number of database-level conformance checks, including:

- Fields have nonpermissible missing values (PCORnet check 1.07)[8]
- Tables have primary key definition errors (e.g., checking whether patient identifiers in the demographics table are unique) (PCORnet check 1.05)[8]

Relational databases also include constraints applied at both the table and field levels. These constraints define the type, quantity, or value of data such that entries conform to prespecified standards (e.g., a data dictionary or codebook). Such constraints may be defined by the style and length of specific code sets (e.g., ICD-9 or ICD-10), whether data elements are known to be numeric or letter-based, or according to known ranges (e.g., nonnegative values). Some PCORnet and Sentinel examples include:

- Fields do not conform to data model specifications for data type, length, or name (PCORnet check 1.04)[8]
- Patient identifier variable type does not conform to specifications (Sentinel check COD_1_01_00-0_112)[25]

* Primary keys specify columns that are both defined as not null and can be used to identify rows in a table (e.g., unique record IDs). Foreign keys specify columns where the values reference those of the primary key in a different table but can accommodate null values. Uniqueness requires values in a column be unique, and not null prevents null values in a given column.[13]

- Birth date variable length does not conform to specifications (Sentinel check DEM_1_02_00-0_113)[25]
- Patient identifier value contains special characters (Sentinel check COD_1_01_00-0_125)[25]
- Patient identifier value contains leading spaces (Sentinel check COD_1_01_00-0_122)[25]

A final form of conformance check regards whether values derived from the computation of other data elements are correct according to the specified function. Computational conformance checks only verify that computations are accurate (e.g., that a BMI calculated from a weight of 250 lbs and a height of 48 inches is 76.3), not that the resulting values are plausible. For example, Kaiser Permanente's CESR applies a conformance check that verifies that the length of stay for inpatient hospital stays (determined by subtracting the discharge date from the admission date) is computed accurately.[11,20] The actual check is as follows:

- For IP stays only, compute LOS (adate-ddate) (CESR check, check ID unavailable)[11,20*]

Completeness Verification Checks for EHRs

Verification checks for data completeness assess "the frequencies of data attributes present in a dataset without reference to data values."[2] Although RWD can be expected to have varying rates of missingness, it is necessary to understand the root cause to allow for adjustments in the data accrual or final analyses to improve reliability.[26] Because EHR systems are designed to support clinical care and workflows, data could be missing from a research perspective but complete from a care perspective. Assessing completeness to characterize differences in missing data between sources can further inform research design. If data aggregation populates missing data from an original source with data from an alternate source, characterizing the completeness of only the output may result in inaccurate representations of data reliability.

To evaluate what data may be missing, completeness checks often use key variables and a defined denominator (e.g., total patients in a database, or patients with a certain feature). While key variables largely depend on the research question, some variables are consistent across EHR databases (e.g., identification and demographic variables). Anticipating and understanding how denominators may change is an important consideration. For example, denominators are affected when a patient dies, leaves the health system, or does not follow up, as well as when a health system loses access to an external data source. While a certain amount of data loss is acceptable, attention to the stability of denominators over time can raise awareness of potential reliability faults. Completeness checks to investigate denominator stability may be designed in units of time or persons. However, designing and implementing data stability checks based on time lacks standardization. As an example, potential

* IP: Inpatient; LOS: Length of Stay; Adate: Admitted Date; Ddate: Discharge Date

stability checks might leverage artificial intelligence or machine learning to identify patterns of missingness.

❓ Plausibility Verification Checks for EHRs

Verification checks for plausibility assess the "believability or truthfulness of data values."[2] For EHR data, plausibility checks for uniqueness can identify situations in which there is conflicting information within a patient record or at the site level. These duplication errors can occur within relational databases when combined objects overlap or when data extraction errors occur.[2] For example, CESR checks whether patient records reflect duplicate admissions to the same facility, which would not be plausible.[11,20]

The check is as follows:

- **Classify how many records:**
  1 = no duplicate admission to same facility,
  2+ = duplicate admission (CESR check, check ID unavailable)[11,20]

Chapter 3: Special Considerations for Verification Checks to Assess PGHD Reliability

Background

The Office of the National Coordinator for Health Information Technology (ONC) defines patient-generated health data as "health-related data created, recorded, or gathered by or from patients (or family members or other caregivers) to help address a health concern."[27] However, as these types of data are also collected for general wellness purposes and potentially throughout a person's life as he or she moves in and out of specific disease states, it is increasingly common to term these data "person-generated." Therefore, we propose to define person-generated health data (PGHD) as wellness and/or health-related data created, recorded, or gathered by individuals for themselves (or by family members or others who care for an individual).

Person-Generated Health Data (PGHD)

Wellness and/or health-related data created, recorded, or gathered by individuals for themselves (or by family members or others who care for an individual)

PGHD reflect events and interactions that occur as part of an individual's everyday life, and therefore recording this information may be more immediately meaningful to patients than information traditionally collected as part of clinical trials. However, PGHD is an emerging field with an ever-increasing amount of data collected through an ever-increasing number of devices, apps, and websites, with few standard data definitions or formats. To ensure the reliability of PGHD for research, data must be assessed and appropriately characterized. The ability to link between data sources and data types is particularly important with PGHD, as this type of data often will serve as an important supplement to clinical data from EHRs or administrative claims data. Baseline checks on the data elements needed to accurately connect patient data to other sources of PGHD or other types of data (e.g., EHR and claims) are crucial.

The use of PGHD to support regulatory decision making is more recent than the use of EHRs and claims data, and specifying verification checks to assess reliability is complicated by the multiple potential sources and methods of collection. During the workshop, the PGHD breakout session used a modified version of PCORnet Data Quality checks to spur discussion of data reliability assessments (Appendix J).[8] However, rather than identifying certain checks that may always be applicable to reliability assessments, the group considered the concepts that drove each check and used the checklist as a framework to discuss how to assess the reliability of PGHD. Thus, using the PCORnet data checks as a starting point, we assembled a list of examples to guide a conversation on developing more specific baseline verification checks on conformance, completeness, and plausibility for PGHD as fit-for-use RWD, similar to those discussed in the previous chapter on checks used with EHR. These examples should continue to be refined and tested with real-world case studies to develop consensus standards on a minimum set of

verification checks that can help support assessments of data reliability. Pilot studies could provide a useful roadmap of the end-to-end process of selecting and ensuring validation of a tool to collect PGHD, linking PGHD to any additional required RWD, and verifying and validating the resulting real-world dataset as fit for use. These studies should make use of a variety of types of PGHD, as well as different regulatory and clinical contexts.

Types of Person-Generated Health Data

In the Framework for FDA's Real-World Evidence Program released in December 2018, FDA provided examples of RWD, explicitly including "patient-generated data, including from in-home-use settings; and data gathered from other sources that can inform on health status, such as mobile devices."[1] Because PGHD are generally collected during an individual's normal routines, some types of PGHD may be considered RWD even when collected specifically for a clinical study, as they are "relating to patient health status" and are being "routinely collected" in the daily life course. It may be helpful to consider a spectrum of data that includes traditional prospective clinical trial data on one end and observational retrospective data on the other, and the many types of PGHD (see Figure 4) collected for a variety of purposes fitting in at different intervals along that space. The verification checks described in this paper are applicable across this spectrum, whether PGHD were primarily collected for personal or clinical use or as part of a clinical study.

FDA has shown considerable interest in PGHD, both as a source of RWD and collected as part of a clinical trial. Table 1 shows examples of work that FDA and its partners have done in this space.

Table 1: Examples of FDA-Supported Projects Advancing the Use of PGHD.

EXAMPLE	DESCRIPTION
MyStudies App (FDA)	FDA recently released the MyStudies App which is "designed to facilitate the input of real-world data directly by patients which can be linked to electronic health data supporting traditional clinical trials, pragmatic trials, observational studies and registries."[28]
Patient-Focused Drug Development program (FDA)	A program focused on incorporating the patient's voice into the drug development and decision-making process, including a focus on selecting, developing, and modifying fit-for-purpose clinical outcome assessments to measure outcomes of importance to patients.[29]
Mobile Clinical Trials Program (CTTI)	The Clinical Trials Transformation Initiative (CTTI), a public-private partnership that includes FDA, the Centers for Medicare & Medicaid Services, and the National Institutes of Health, also recently completed work on a mobile clinical trials program, that looked at how researchers can choose mobile technologies, develop novel end points, and run decentralized clinical trials.[30]
Characteristics of PGHD from a novel source (PatientsLikeMe)	Through a research collaboration agreement, PatientsLikeMe and FDA conducted several research projects to explore and analyze the characteristics of PGHD from a novel source to inform regulatory review activities, including an evaluative study of curation and coding practices of patient verbatim reports of treatment side effects.[31]
mHealth Action Plan	The Duke-Margolis Center, under a cooperative agreement with FDA's Center for Devices and Radiological Health, released a white paper in 2017 with recommendations on how to make mHealth data a reliable source of RWD.[32]

Figure 4. The Four Types of PGHD*

The different types of PGHD can be categorized into four groups: person-reported data, task-based measures, active sensor data, and passive sensor data.[32] Each of these types could be used individually or in combination with each other or other types of RWD or clinical trial data, allowing for a broad range of research studies. Sensors would include both consumer devices and medical devices that can collect personal data.

PERSON-REPORTED DATA

Data reported manually by the person themselves (or their caregiver if the person is unable to enter the data)

Person-reported data may include responses to questionnaires, symptom and behavior tracking, or other means of collecting person-reported outcomes. Historically, person-reported data have been captured through paper-based and web-based surveys, phone calls, and so forth. However, such data can also be collected through mHealth apps/websites and even through applications on wearable devices during the course of everyday life. These data could be used to collect patient-reported outcomes (PROs, also known as ePROs when collected electronically)—measurements based on reports that come directly from patients about the status of patients' health conditions without interpretation by anyone else—or observer-reported outcomes (ObsROs)—measurements based on reports of observable signs, events, or behaviors related to a patient's health condition by someone other than the patient or a health professional.[33] A daily pain diary is an example of an ePRO collected for the purposes of regulatory decision making is a daily pain diary.

TASK-BASED MEASURES

Objective measurement of a person's mental and/or physical ability to perform a test consisting of a defined task or set of tasks. Task-based measures require cooperation and motivation.

Task-based measures may include physical functioning tests (e.g., 6-minute walk test) or cognitive functioning tests (e.g., digit symbol substitution test) performed by the patient or consumer. Some of these measures (also known as performance outcomes or PerfOs) historically are captured in a clinical setting with appropriate clinical or task procedure validation, but they may be captured in real-world settings if appropriate instructions and methods are adequately described to the individual performing the task and there is confirmation that the task is performed as directed. Task-based measures can be collected through remote sensors and/or mobile apps that may use sensors within the smartphone, and there are already many examples of apps with this functionality, including the "Active Tasks" that are part of Apple's ResearchKit, which collect data on measures of spatial memory, tapping speed, reaction time, and more.[34]

ACTIVE SENSOR DATA

Measurement of a person's daily activities, mental state, or physiological status that requires an activation step (e.g., stepping on a scale, glucose self-measurement)

Active sensors require an activation step for a measurement to be taken. Active sensor data differ from task-based measures because, while the individual must perform an action to collect data, the ability to perform that action is unrelated to the type of data being collected. In contrast, task-based measures ask individuals to perform tasks for the purpose of collecting data on how well they are able to complete the task. For example, the ability to easily step on a scale is unrelated to the resulting measurement of the individual's weight.

PASSIVE SENSOR DATA

Measurement of a person's daily activities, mental state, or physiological status that does not interrupt the person's normal activities. (Note that measurement of daily activities reflects what an individual actually does in their daily life, not a measure of what they are capable of or comfortable doing.)

Sensors such as wearable and remote sensors (both consumer-grade and FDA-approved/cleared), sensors on mobile devices, and tools that monitor behavior (e.g., analyses of changes in social media habits) can passively collect information about people's daily lives. This can include measures such as activity level, heart rate, and sleep patterns. Passive sensing has the benefit of being "invisible" because it does not require active interaction and therefore is less disruptive of normal routines. In addition, these tools have the ability to capture data during times (or over lengths of time) and locations that may allow the data to be more representative of an individual's state.

* Figure modified from 2017 Duke-Margolis mHealth Action Plan.[32]

Context for Assessing the Reliability of PGHD

While the use of patient experience and medical device data has been detailed in clinical literature, the reliability of other types of PGHD can be difficult to characterize. For example, when considering the reliability of PGHD, sensor data taken during daily activities or an actual disease event may be different than data recorded in controlled, simulated settings. For example, data collected during a naturally occurring seizure may give different information than data collected during seizures induced in clinical testing. Surveys given more frequently or in response to an event detected through a sensor may yield more useful information than reports taken at the clinic that may require longer recall periods.

However, the reliability of the data may also depend on factors such as a person's understanding of any medical terms used in surveys or instructions and the quality of the sensors being used. Characterizing the data will require a deep understanding of how people are using the data collection tools in their daily lives, which may differ from researchers' assumptions. For example, Biogen and PatientsLikeMe gave activity trackers to 200 people with multiple sclerosis to monitor and manage their condition for a study.[35] Some patients used the trackers as passive sensing devices, but others (after noticing personal patterns) used the devices to self-limit their activity in order to manage subsequent symptoms.[36] Conformance and consistency can also be a concern with PGHD, because much of this type of data remains unstandardized and has little transparency around transformations the data may have undergone, particularly with consumer devices. While some organizations have started to develop data dictionaries and standards, much work is left to be done.[37,38] Completeness of data may be an issue when patients fail to consistently wear, charge, or sync a device, or miss opportunities to record their data. These missed opportunities may differ according to patient outcomes (e.g., good health vs. poor health), time pressures (e.g., employment), or other factors that may bias results and interpretation.

Therefore, when choosing a PGHD data source, a researcher first must take care to assess the tool(s) used for collection, which includes both the outcome instrument and the platform, device, or software program collecting the data. Several groups have released recommendations on the verification and validation of tools related to PGHD collection. The CTTI Mobile Clinical Trials Program has published recommendations on verification of PGHD collection tools as well as novel end points.[39,40] Other efforts to characterize and standardize PGHD collection tools include the FDA Clinical Outcome Assessment Qualification Program; the Consumer Technology Association's Health, Fitness and Wellness Subcommittee; and the Critical Path Institute's Electronic Patient-Reported Outcome Consortium.[41-43]

Second, the researcher must assess the reliability of the person-generated dataset itself, which is the focus of this paper. Creating standards or best practices for performing this assessment is a challenge due to the enormous range of possible data that could be collected, the lack of experience and standards in assessing this continuously growing ecosystem of data types, and because this type of data generally must be linked to other sources of clinical data to ensure that researchers have access to a detailed enough clinical picture of the study subjects.

Other challenges include the potential for software updates that affect how sensor data are analyzed over time, which can be a particular problem when the lack of on-device storage means that the raw sensor data are not stored and made available for analysis, and instead only transformed or processed data are available. For example, consumer activity trackers frequently do not store the actual accelerometer sensor data, which measure changes in gravitational force. Instead, the raw data are analyzed in real time and converted into "steps" and other activity measures, which are stored on the device.

Finally, there are differing views on the rights of people to control the access and use of some types of PGHD, especially data collected through consumer apps and sensors. National and state governments are beginning to pass and implement various laws to protect this sort of data, which could affect information on data provenance.[44,45]

Using Verification Checks to Assess PGHD Reliability

Unlike EHRs and administrative claims data, few common data models or standard common data dictionaries for PGHD exist because of the heterogeneity of the types of data that are recorded and the methods for collecting those data. Therefore, researchers and regulators developing verification checks will be required to prespecify a data model, with key variables and metadata also pre-identified. These key variables will generally be specific to a particular study, including exposures, covariates, outcomes, variables required for linking person-specific data to other data sources, etc. However, some examples of key variables and metadata are likely to be fairly common across studies, including person-identifier data allowing participants to be tracked over time, data source identifiers (including applicable sensor or hardware identifiers), firmware and software versioning information, and data timestamps. The data model should include the expected data elements with data type, frequency, data length restrictions, name, and definitions. For the purposes of verification checks, we are primarily concerned with ensuring that this information is included in the selection of key variables and metadata, so that users and regulators can be assured that the data required for the selected methods and for showing appropriate generalizability are present, are in a usable format, and are accurate. Once the data model is specified and the key variables and metadata have been identified, verification checks can be formulated.

Conformance Verification Checks for PGHD

Verification checks for data conformance assess the structure of the data and how compliant the data are with internal relational, formatting, or computational definitions and standards,[2] as set out in a prespecified data model.

Examples of conformance checks that could be used with structured PGHD include the following:

1. **Are the prespecified required elements of the data model present? For example, are the required fields or tables present?**

   Note that this is a structural requirement only; whether there are data in those fields or tables is considered in later checks.

2. **Do the fields conform to the prespecified data model specifications for data type, length, or name?**

   If not, the dataset can be cleaned and/or transformed to improve conformance; however, these actions should neither exclude data nor degrade data integrity. Adequate data maintenance preserves end users' ability to audit their mapping and transformations, processes that should be prespecified and justified.

3. **Do the data assign more than one person identifier to the variable? More than one timestamp?**

   For verification checks like these, it will be useful to assume that some data will not meet the standard. However, some amount of nonconforming data will still result in a usable dataset. In these cases, it is helpful to prespecify a threshold for which the data could still be characterized overall as fit for use.

4. **Do key fields have nonpermissible missing values?**

   If the PGHD collection software is thought to be programmed to prevent certain values or prevent data from being collected without ensuring specific fields are completed, then these missing values suggest that either the prespecified data model expectations were in error or that the software was programmed incorrectly, which could suggest additional challenges with the data.

5. **Are the variables derived from computations correct according to the specified function?**

   Computational conformance checks only verify that computations are accurate according to the specified function and the data used, not that the resulting values are plausible. Therefore, a BMI of 76.3 calculated from a weight of 250 lbs and a height of 48 inches would meet a conformance check, but may be flagged by a plausibility check (see later section).

Completeness Verification Checks for PGHD

Verification checks for data completeness assess "the frequencies of data attributes present in a dataset without reference to data values."[2] These checks measure how a data environment aligns with expectations and analysis needs. For example, for a survey given weekly, the expectation would be that data would be available for each 7-day period. Continuously collected data, however, may need to be considered somewhat differently.

Examples of completeness checks for data verification could include:

1. **How many records have missing or unknown values for prespecified key variables?**

   The acceptable range of missingness for key variables will depend on the clinical context, the regulatory decision, and the importance of the particular variable to the overall analysis. Key variables may include data on exposure, covariates, and outcomes, but will also need to include data that are required to analyze the data to generate RWE. For example, completeness checks for verification will record how much data are missing, but not whether the data are "missing at random" or "not at random." In many cases, this information is required to analyze the data appropriately, so variables or metadata that will help researchers make those determinations should also be deemed key variables and completeness data collected.

2. **How many records have missing or unknown values for prespecified key metadata?**

   Examples of metadata for PGHD include timestamps, time zone data, person identifiers, date of sign up, date of last activity, and versioning information on any hardware, operating system, firmware, software, and APIs that are used in the collection and sharing of the PGHD. Other metadata that may be useful include whether data were manually entered, were calculated from other data, or were directly entered from a sensor (in which case a sensor identifier should be included).

3. **Does each study subject have a prespecified minimum number of interactions or uses with the PGHD collection tool that resulted in data collection?**

   PGHD are unique in that the researcher will need to consider how "completeness" balances with "expected sampling frequency" and/or "expected use." The amount of data that a researcher should expect from a device that collects data continuously (e.g., an activity tracker or a continuous glucose monitor) will be different than the number of times a person may record changes in medication or answer survey questions.

4. **Does data completeness change over time?**

   Changes in data completeness over time may indicate multiple issues. For example, the collection device may be broken; the software driving the collection device may have changed; the device may not be communicating properly with the data collection source; or perhaps the study subject stopped using the device. Each of these issues has implications for the analysis, so this information is critical to assessing the reliability of the dataset. When possible, metadata that could indicate the reason for any changes should be identified and included as a key

variable in the verification checks. This may require data from other sources, such as EHRs, in which case data required to link to those sources should be identified and included.

Plausibility Verification Checks for PGHD

Verification checks for plausibility assess the "believability or truthfulness of data values."[2]

Examples of plausibility checks for data verification could include:

1. **Within prespecified fields, how much of the data are in the highest and lowest ranges of biological plausibility?**

   This kind of check may not be possible for all of the data, but for sensor data such as heart rate and step counts, and patient-reported data such as age, height, and weight, most data should be in the middle of the normal range for the population of interest. Excessive data points in the highest and lowest ranges may indicate that the sensors were not working as expected or that the questions study subjects were answering were unclear. If the data are accurate, deviation from expected values could have implications for the generalizability of the data.

2. **Are the median patient-reported or sensor data values for selected fields statistical or clinical outliers from a prespecified range?**

   Similar to the verification check above, this check allows a researcher to know if there are accuracy or generalizability concerns with a particular dataset.

3. **Do study subjects have illogical date relationships?**

   Illogical date relationships are another way for researchers to gauge whether their datasets have reliability challenges. Examples of illogical date relationships in PGHD include interactions recorded as occurring before the sign-up date (or even birth date), a sensor capturing activity before it was activated, and an end time for an activity that occurs before the start time.

4. **For each subject, were the number of interactions or the amount of data collected over a certain time frame significantly more than expected? Were data collected more often than indicated in the protocol?**

   These types of checks assess whether the amount of data collected is significantly more than expected (e.g., 1-minute heart rate was collected more than 60 times in an hour, or a subject answered a survey 20 times when it was supposed to be presented only six times). This situation differs from when the data collected are significantly less than expected, which is assessed through completeness checks.

APPENDIX A. WORKSHOP PARTICIPANTS

Principles for Developing Quality Checks to Assess Fit-for-Use RWD

Duke-Robert J. Margolis, MD, Center for Health Policy
1201 Pennsylvania Ave, NW, Suite 500 • Washington, DC 20004
May 28, 2019

Participant List

Aylin Altan
OptumLabs

Marc Berger
ISPOR

Clair Blacketer
Janssen Research and Development

David Blaser
PatientsLikeMe

Jeffrey Brown
Harvard Pilgrim Health Care Institute

Sailu Challapalli
Flatiron Health

Andy Coravos
Elektra Labs

Brande Ellis Yaist
Eli Lilly and Company

Shannon Ferrante
GlaxoSmithKline

Rachael Fleurence
NESTcc

Marni Hall
IQVIA

Kristijan Kahler
Novartis

Christina Mack
IQVIA

Vandana Menon
OM1

Claire Meunier
Evidation

Anne-Marie Meyer
F. Hoffmann-La Roche

Rebecca Miksad
Flatiron Health

Marko Mychaskiw
Teva Pharmaceuticals

Andrew Norden
COTA

Sally Okun
PatientsLikeMe

Marc Overhage
Cerner Corporation

Laura Qualls
Duke Clinical Research Institute

Ernesto Ramirez
Evidation

Jeremy Rassen
Aetion

Stephanie Reisinger
Veradigm Life Sciences, Allscripts

Sam Roosz
Datavant

Patrick Ryan
Janssen Research and Development

Tim Sampson
UCB

Edgar Simard
GlaxoSmithKline

Jason Simeone
Evidera

Vince Willey
HealthCore

Andrew Williams
Tufts Clinical and Translational Science Institute

Lisa Wruck
Duke Clinical Research Institute

Dongmu Zhang
AbbVie

Duke-Robert J. Margolis, MD, Center for Health Policy

Adam Aten
Duke-Robert J. Margolis, MD, Center for Health Policy

Gregory Daniel
Duke-Robert J. Margolis, MD, Center for Health Policy

Adam Kroetsch
Duke-Robert J. Margolis, MD, Center for Health Policy

Nirosha Mahendraratnam Lederer
Duke-Robert J. Margolis, MD, Center for Health Policy

Kerra Mercon
Duke-Robert J. Margolis, MD, Center for Health Policy

Morgan Romine
Duke-Robert J. Margolis, MD, Center for Health Policy

Christina Silcox
Duke-Robert J. Margolis, MD, Center for Health Policy

Observers

Michael Blum
US Food and Drug Administration

Jacqueline Corrigan-Curay
US Food and Drug Administration

David Martin
US Food and Drug Administration

Amanda Wagner Gee
National Academies of Sciences, Engineering, and Medicine

APPENDIX B. GLOSSARY

Because many of the terms used in this paper have different definitions based on the user and discipline, we compiled and developed a list of standard definitions from the perspective of assessing fit-for-use RWD for regulatory decision making. When available, we used FDA definitions and supplemented them with information from the Duke-Margolis white paper *Characterizing RWD Quality and Relevancy for Regulatory Purposes*, a literature review, and feedback from the RWD Quality Working Group.

Data Accrual: The process by which data are collected and aggregated. This includes data provenance.

Data Lineage: The history of all data transformations (e.g., recoding or modifying variables).

Data Model: Standardizes and specifies which data elements will be stored, how they will be stored, and how they are related.[46]

Data Provenance: The origin of the data, sometimes including a chronological record of data custodians and transformations.

Data Relevancy: Assessment of whether the data adequately address the applicable regulatory question or requirement, in part or in whole. This assessment includes whether the data capture relevant information on exposures, outcomes, and covariates and whether the data are generalizable.[1,5]

Data Reliability: Considers whether the data adequately represent the underlying medical concepts they are intended to represent. A broad concept that encompasses data accrual and data quality control (data assurance).[1]

Fit-for-Use RWD: Meets a standard for which data can be successfully used to inform regulatory decision making. FDA asserts that fit-for-use data are both reliable and relevant. In the draft RWE Framework, FDA seems to use these terms interchangeably.

Quality Assurance: Consists of proactive and retrospective activities undertaken to evaluate whether prespecified requirements are fulfilled.[6]

Quality Control: Consists of the steps taken during data curation to ensure that the data meet prespecified standards and are reproducible.[6]

Raw Data: Data in their original form or as collected, prior to any curation.

Real-World Data (RWD): Data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources.[1]

Real-World Evidence (RWE): Clinical evidence about the usage and potential benefits or risks of a medical product derived from analysis of RWD.[1]

Research-Ready Database: A database containing raw RWD that have undergone some data curation processes. Data from a research-ready database can be extracted, and potentially further transformed, into RWD that are fit for use.

Validation: Alignment of data values with respect to relevant external benchmark (e.g., a gold standard).[2]

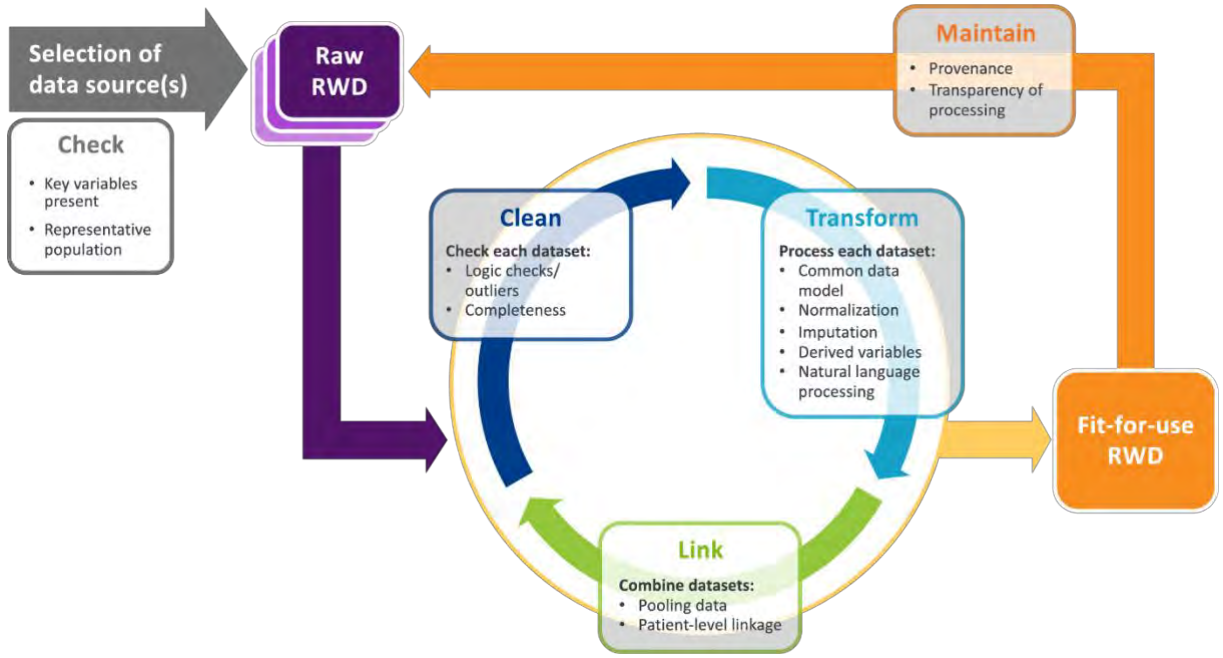
Verification: Assessment of how a data element or variable matches expectations with respect to metadata constraints, system assumptions, and local knowledge.[2]

APPENDIX C. RWD CURATION

RWD collection can be either primary or secondary. Primary RWD are collected for the purpose of answering the specific research question (e.g., a postmarket registry) and generally have varying degrees of quality control and data curation as part of the process of data collection and integration. In contrast, secondary RWD are collected for a different purpose but can be used to answer the research question. For example, claims data are primarily collected for billing, and EHR data are primarily collected to manage clinical workflows as well as billing and reimbursement. Both types of data can be used as a source of RWD for regulatory decision making. Because the original intent is unrelated to research purposes, secondary data in their raw form are often not "research-ready" and require substantial data curation. Even after data curation, not all "research-ready" databases can be used to create a fit-for-use dataset that can be analyzed to generate evidence for regulatory decision making.

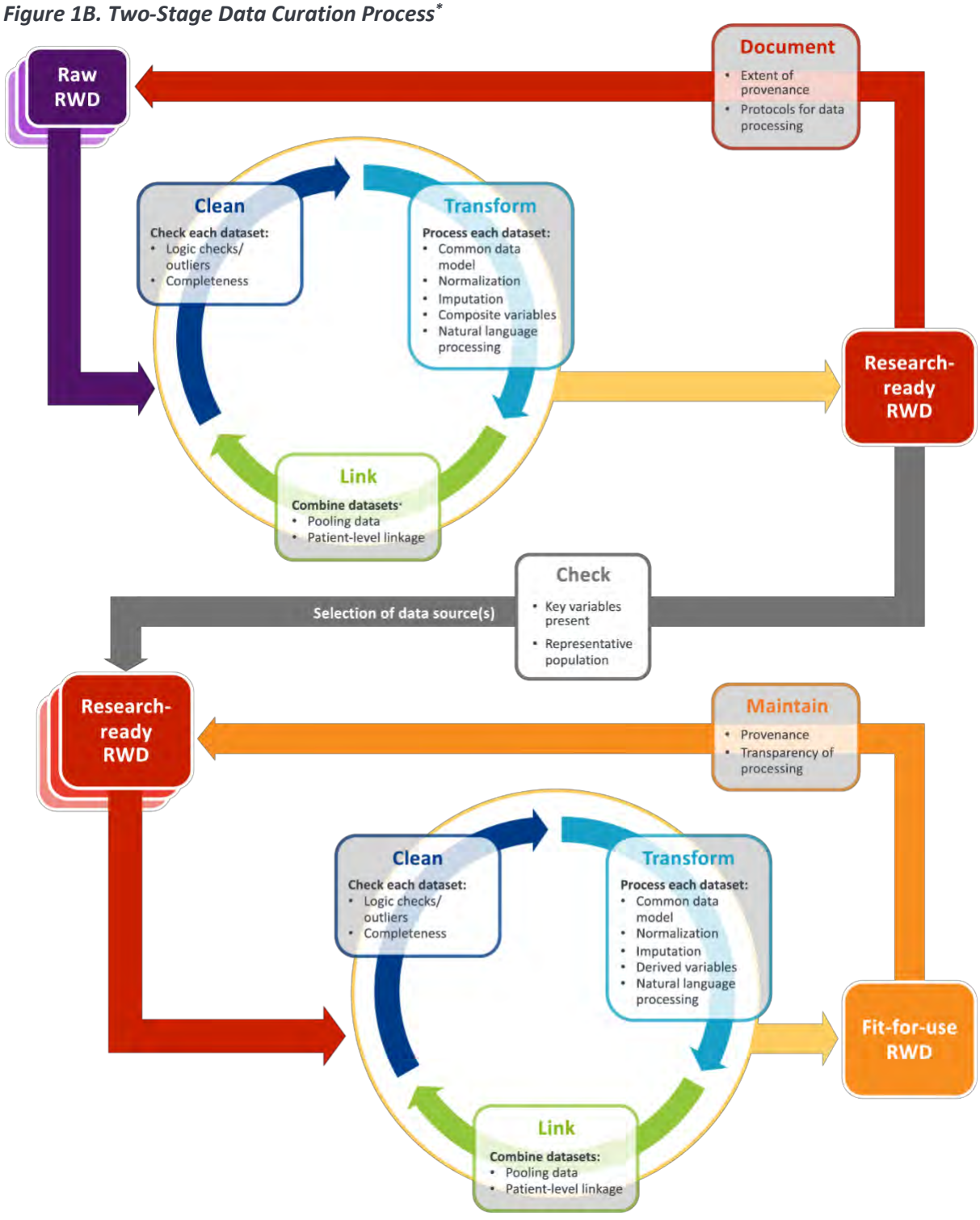
Understanding the RWD curation process is necessary for characterizing and assessing RWD as fit for use. In a 2018 whitepaper, "Characterizing RWD Quality and Relevancy for Regulatory Purposes," we detailed how data curation can be either a singular process or a multistage process (Figures 1A and 1B).[3] The same steps occur regardless of whether it is a singular or multistage process, but the stakeholders involved may differ at various points. In a singular process (Figure 1A), study investigators start with raw RWD and are responsible for all the curation steps required to make the data fit for use. In a multistage process (Figure 1B), a third party typically acquires the raw data from a variety of RWD sources and normalizes the data into databases that could be used for general research purposes (i.e., entering the data into a data model and/or applying data standards), otherwise referred to as research-ready RWD. With additional curation steps, selected data from these databases can be further curated by researchers into a fit-for-use real-world dataset to answer a specific study question, which represents the second stage of data curation (and multiple data sources can comprise a fit-for-use dataset). Whether a singular or multistage curation process, there may be cycles of data cleaning steps (e.g., logic checks, assessments of data missingness), data transformations (e.g., data model mapping, normalizing data values), and data linkages (e.g., combining data from different sources). Data curation steps are not singular but rather exist on a continuum. These steps happen multiple times, often in tandem, as raw data are translated to research-ready data, and then to fit-for-use data. It is vital to document and explain any changes or exclusions in the underlying data during these processes. The rules should be explicitly stated and the magnitude of the impact should be reported (e.g., the percentage of records that were changed or excluded).

Figure 1A. Single-Stage Data Curation Process*



* Figure modified from 2018 Duke-Margolis *Characterizing RWD Quality and Relevancy for Regulatory Purposes*.[3]

Figure 1B. Two-Stage Data Curation Process[*]



[*] Figure modified from 2018 Duke-Margolis *Characterizing RWD Quality and Relevancy for Regulatory Purposes*.[3]

APPENDIX D. APPROACHES TO RWD CURATION PROCESSES

Identifying Data Curation Best Practices as Quality Control to Support RWD Reliability

There are several approaches used to curate data. These include human abstraction methods and automated extraction techniques (e.g., natural language processing) to extract information from the unstructured component of medical charts as well as strategies for standardizing data into data models (e.g., data transformations), linkage of data sources, and study-specific curation practices for developing a fit-for-use dataset. Identifying data curation best practices that serve as the gold standard for quality control is one way to evaluate fitness for use. In January 2019, under a cooperative agreement with FDA, the Robert J. Margolis, MD, Center for Health Policy held a private workshop entitled "Unpacking Real-World Data Curation: Principles and Best Practices to Support Transparency and Quality" to better understand RWD fitness for use from diverse stakeholder perspectives. During this meeting, a number of key data reliability issues were identified across both data curation stages (Appendix C, Figure 1B), but the discussion did not clearly identify any common practices or comparative advantages of specific curation tools. Given the diversity of source data, evolving technology landscape, and other key considerations such as therapeutic context, it is unlikely that curation best practices can be identified.

Using Data Curation Documentation to Support RWD Reliability

Documentation is the process of recording how and why data are transformed (e.g., recoding or modifying variables). Transformations can be recorded using different mechanisms including, but not limited to, metadata, code, or text.[6] However, there is a tension between transparency and interpretability of the documentation submitted describing data curation practices. It is important to document data reliability issues (e.g., missingness, acceptable data value range, and correct linkages of data points to patients) and to document the impact of dropping data elements. Furthermore, documenting the verification and validation of tools for PGHD collection may be desired, and such information can be included in protocols and statistical analysis plans. However, submitting a large amount of information can affect the interpretability of this information. For example, providing the underlying source code of the transformations, while potentially useful, may not be practical for FDA reviewers to efficiently evaluate.

Using Checks as Quality Assurance to Support RWD Reliability

Given current difficulties with identifying a common set of data curation practices that could be systematically evaluated, it was suggested that a set of quality assurance practices, specifically checks, could help regulators determine a dataset's fitness for use. Checks can be used to assess different aspects of RWD fitness for use, including reliability. These checks would use statistical measures or standardized summary documents to indicate whether data transformations "worked" to address data reliability needs. It is important to remember that because data curation is a dynamic process that sits on a continuum, checks should also be continuously assessed and may need to be adapted based on observed data reliability issues.

APPENDIX E. VERIFICATION CHECKS TO ASSESS DATA RELIABILITY

Table 1: Complete List of Verification Checks to Assess Data Reliability Cited-in-Text

DATA NETWORK	DESCRIPTION	CHECK ID	REF
CONFORMANCE VERIFICATION CHECKS			
PCORnet	Fields have non-permissible missing values	DC 1.07	8
PCORnet	Tables have primary key definition errors	DC 1.05	8
PCORnet	Fields do not conform to data model specifications for data type, length, or name.	DC 1.04	8
Sentinel	PatID variable type does not conform to specifications	COD_1_01_00-0_112	25
Sentinel	Birth_Date variable length does not conform to specifications	DEM_1_02_00-0_113	25
Sentinel	PatID value contains special characters	COD_1_01_00-0_125	25
Sentinel	PatID value contains leading spaces	COD_1_01_00-0_122	25
CESR	For IP stays only, compute LOS (adate-ddate)	Not Available	11,20
COMPLETENESS VERIFICATION CHECKS			
PCORnet	Less than 50% of patients with encounters have DIAGNOSIS records	DC 3.04	8
PCORnet	The average number of diagnoses records with known diagnosis types per encounter is below threshold [1.0 for ambulatory (AV), inpatient (IP), emergency department (ED), or ED to inpatient (EI) encounters]	DC 3.01	8
PLAUSIBILITY VERIFICATION CHECKS			
CESR	Classify how many records: 1 = no duplicate admission to same facility, 2+ = duplicate admission	Not Available	11,20
MURDOCK	HEIGHT is not between 36 and 84	Not Available	11,18
PCORnet	More than 10% of records fall into the lowest or highest categories of age, height, weight, diastolic blood pressure, systolic blood pressure, or dispensed days supply	DC 2.02	8
PCORnet	More than 5% of records have future dates	DC 2.01	8
CESR	Count number of observations by year across all years of data	Not Available	11,20
Sentinel	ADate value occurs after DDate value	ENC_2_03_00-0_226	25

APPENDIX F. DATA LINKAGE

Ensuring that all information for a given patient is aggregated is an important part of the data curation process. Data linkage is the process of identifying and matching records that belong to the same entity and is crucial for limiting the duplication of records.[47] Linking relies on identifying data with overlapping or common identifiers between multiple sources or within a single dataset. Because the availability of identifying information may differ among data sources, it is important to check whether linking occurs as intended. In particular, ensuring that patients are uniquely differentiated is critical to supporting data reliability. In the United States, no universal unique patient identifier connects patients across health systems, and data are typically, although not always, deidentified before being used for research. Therefore, when working with RWD from multiple sources, it is often necessary to develop methods to identify and link patient records based on demographic, geographical, or historical information.[21]

APPENDIX G. ELECTRONIC HEALTH RECORDS AND MEANINGFUL USE

Electronic Health Record Definition

As defined in section 300jj of Title 42 of the U.S. Code, an electronic health record is an "electronic record of health-related information on an individual that includes patient demographic and clinical health information, such as medical history and problem lists; and has the capacity—(i) to provide clinical decision support; (ii) to support physician order entry; (iii) to capture and query information relevant to health care quality; and (iv) to exchange electronic health information with, and integrate such information from other sources."[23]

Meaningful Use of Electronic Health Records

According to the Health Information Technology for Economic and Clinical Health (HITECH) Act, meaningful use of EHRs incudes "(1) Improving quality, safety, efficiency, and reducing health disparities; (2) Engage patients and families in their health; (3) Improve care coordination; (4) Improve population and public health; and (5) Ensure adequate privacy and security protection for personal health information."[23]

APPENDIX H. EHR BREAKOUT SESSION TOOL

Perspective

The following table is modified based on the PCORnet Quality Checks v6. The threshold values have been replaced with an "x" in order to focus discussion on the check and not the value of the threshold. The following checks have been developed specifically for the PCORnet Common Data Model for a Distributed Data Network. For this meeting, we ask that everyone review the checks from the perspective of your OWN ORGANIZATION'S data model (or standard way of organizing and structuring data). We recognize that some organizations may feed data into a Distributed Data Network, but we are interested in the perspective of checks for YOUR ORGANIZATION.

Definitions

Primary Key: One or more columns in a table that distinguishes one row from another row in a table[48]

Foreign Key: One or more columns in a table that reference a primary key in another table[48]

Orphan: A record where a foreign key value has a non-existent primary key value[48]

Conformance Quality Checks

PRINCIPLE	NUMBER	DESCRIPTION	COMMENTS
DATA MODEL CONFORMANCE (Assessment of the structure of the data and how compliant the data are with internal or external formatting, relational, or computational definitions/ standards)	**1**	Fields do not conform to data model specifications for data type, length, or name	
	2	More than x% of ICD, IPT, LOINC, RXCUI, NDC codes do not conform to the expected length or content	
	3	Fields contain values outside of data model specifications	
	4	Fields have non-permissible missing values	
	5	More than x% of encounters are assigned to more than one patient identification number	
	6	Required fields in a table are not present	
	7	Required tables in the data model are not present	
	8	Expected tables in the data model are not populated	
	9	Tables have primary key definition errors	
	10	Tables contain orphan patient identification numbers	
	11	Tables contain orphan encounter identification numbers	
	12	Tables contain orphan provider identification numbers	
	13	Replication errors between the encounter, procedures, and diagnosis tables (e.g., the encounter identification number in the procedures or diagnosis tables does not match the corresponding variable in the encounters table)	

Completeness Quality Checks

PRINCIPLE	NUMBER	DESCRIPTION	COMMENTS
DATA COMPLETENESS (Evaluates the presence of data by measuring the frequencies of data attributes present in a dataset without reference to data values)	**14**	Less than x% of patients with encounters have diagnosis records	
	15	Less than x% of patients with encounters have procedures records	
	16	Less than x% of quantitative results for tests mapped to LAB_LOINC fully specify the normal range	
	17	The average number of diagnoses records with known diagnosis types per encounter is below the threshold for ambulatory, inpatient, emergency department, or Emergency Department to inpatient encounters	
	18	The average number of procedure records with known procedure rates per encounter is below the thresholds for ambulatory, inpatient, emergency department, or Emergency Department to inpatient encounters	
	19	More than x% of records have missing or unknown values for key variables (e.g., birth date, sex, discharge disposition, discharge date, prescription order date, procedure date, days supply, terminology used to describe clinical observation, standardized code for clinical observations based on set terminology/vocabulary, standardized code for other observations based on set terminology/vocabulary, medication code, medication code type, foreign keys associated with the encounter identification number)	
	20	More than x% of inpatient or emergency department to inpatient encounters with any diagnosis don't have a principal diagnosis	
	21	Encounters, diagnoses, or procedures in an ambulatory, emergency department, emergency department to inpatient, or inpatient setting are less than x% complete three months prior to the current month	
	22	Vital, prescribing, or laboratory records are less than x% complete three months prior to the current month	
	23	Less than x% of prescribing orders are mapped to a RXNORM_CUI which fully specifies the ingredient, strength, and dose form	
	24	Less than x% of laboratory results are mapped to LAB_LOINC	
	25	Less than x% of qualitative results for tests mapped to LAB_LOINC fully specify the source of the specimen and standardized unit for quantitative results	

Plausibility Quality Checks

PRINCIPLE	NUMBER	DESCRIPTION	COMMENTS
DATA PLAUSIBILITY (Assessment of the believability or truthfulness of data values)	**26**	More than x% of records have future dates (dates beyond the day of refresh)	
	27	More than x% of records have extreme values (values that fall in the lowest or highest categories of age, height, weight, diastolic blood pressure, systolic blood pressure, or dispensed days supply)	
	28	More than x% of patients have illogical date relationships (e.g., death date before birth date, dispense date occurs after death date)	
	29	The average number of encounters per visit per patient is greater than "x" number for inpatient (IP), emergency department (ED), or Emergency Department to inpatient (EI) encounters. A high number of encounters per visit may signal potential redundancy or duplication.	
	30	More than x% of results for selected laboratory tests do not have the appropriate specimen source	
	31	Median lab result values for selected tests are statistical or clinical outliers	
	32	The average number of principal diagnoses per encounter is above the threshold for inpatient and Emergency Department to inpatient	
	33	More than x% decrease in the number of records in a Common Data Model table between the previous and current data refresh	
	34	More than x% decrease in the number of patients with diagnosis, procedures, labs, or prescriptions during an ambulatory, emergency department, or inpatient encounter between the previous and current data refresh	
	35	More than x% decrease in the number of records for ICD9 or ICD10 diagnosis or procedure codes or CPT/HCPCS procedure codes between the previous and current data refresh	

APPENDIX I. RELATIONAL DATABASES

Relational databases provide structure and impose constraints on the organization of data. As RWD from EHRs are collected, they are often prepared for analysis within a relational database. Depending on the EHR platform, data may not be natively stored in a relational manner, but may regardless be incorporated into a relational database. Furthermore, data from multiple EHR sources can be combined into a single relational database to produce a research-ready dataset.

The use of relational databases supports data reliability by limiting the opportunity for error and by imposing constraints and defining relationships between data elements. Data in a relational database are organized in tables where each row signifies a unique observation, identifiable by a primary key, and only data directly related to that observation are included.[49] This organization is referred to as normalization, and it ultimately serves to reduce redundancy and promote data integrity as changes are made to the database.

Relational databases also impose constraints on data entry such that the structure of a data model can be maintained. Constraints ensure the integrity of a data model by allowing only expected types of data to be entered into a relational database, and by defining relationships between data elements. For instance, a database may impose a constraint that only allows an encounter ID to be added when an associated patient ID is included.

APPENDIX J. PGHD BREAKOUT SESSION TOOL

Perspective

The following table is modified based on the PCORnet Quality Checks v6. The threshold values have been replaced with an "x" in order to focus discussion on the check and not the value of the threshold. The following checks have been developed specifically for the PCORnet Common Data Model for a Distributed Data Network. For this meeting, we ask that everyone review the checks from the perspective of your OWN ORGANIZATION'S data model (or standard way of organizing and structuring data). We recognize that some organizations may feed data into a Distributed Data Network, but we are interested in the perspective of checks for YOUR ORGANIZATION.

Definitions

Primary Key: One or more columns in a table that distinguishes one row from another row in a table[48]

Foreign Key: One or more columns in a table that reference a primary key in another table[48]

Orphan: A record where a foreign key value has a non-existent primary key value[48]

Legend

Grey Shading: Check is likely not applicable to PGHD and cannot be modified to be applicable.

Conformance Quality Checks

PRINCIPLE	NUMBER	DESCRIPTION	COMMENTS
DATA MODEL CONFORMANCE (Assessment of the structure of the data and how compliant the data are with internal or external formatting, relational, or computational definitions/ standards)	**1**	Fields do not conform to data model specifications for data type, length, or name	
	2	Fields contain values outside of data model specifications	
	3	Fields have non-permissible missing values	
	4	More than x% of the time, patient-reported or sensor data are assigned to more than one patient identification number	
	5	More than x% of ICD, IPT, LOINC, RXCUI, NDC codes do not conform to the expected length or content	
	6	Required fields in a table are not present	
	7	Required tables in the prespecified data model are not present	
	8	Expected tables in the prespecified data model are not populated	
	9	Tables have primary key definition errors	
	10	Tables contain orphan patient identification numbers	
	11	Tables contain orphan encounter identification numbers	
	12	Tables contain orphan provider identification numbers	
	13	Replication errors between tables	

Completeness Quality Checks

PRINCIPLE	NUMBER	DESCRIPTION	COMMENTS
DATA COMPLETENESS (Evaluates the presence of data by measuring the frequencies of data attributes present in a dataset without reference to data values)	**14**	The average number of diagnoses records with known diagnosis types per encounter is below the threshold for ambulatory, inpatient, emergency department, or Emergency Department to inpatient encounters	
	15	The average number of procedure records with known procedure rates per encounter is below the thresholds for ambulatory, inpatient, emergency department, or Emergency Department to inpatient encounters	
	16	More than x% of records have missing or unknown values for key variables	
	17	Less than x% of patients with encounters have diagnosis records	
	18	Less than x% of patients with encounters have procedures records	
	19	More than x% of inpatient or emergency department to inpatient encounters with any diagnosis don't have a principal diagnosis	
	20	Encounters, diagnoses, or procedures in an ambulatory, emergency department, emergency department to inpatient, or inpatient setting are less than x% complete three months prior to the current month	
	21	Less than x% of prescribing orders are mapped to a RXNORM_CUI which fully specifies the ingredient, strength, and dose form	
	22	Less than x% of laboratory results are mapped to LAB_LOINC	
	23	Less than x% of quantitative results for tests mapped to LAB_LOINC fully specify the normal range	
	24	Data are less than x% complete three months prior to the current month	
	25	Less than x% of qualitative results for tests mapped to LAB_LOINC fully specify the source of the specimen and standardized unit for quantitative results	

Plausibility Quality Checks

PRINCIPLE	NUMBER	DESCRIPTION	COMMENTS
DATA PLAUSIBILITY (Assessment of the believability or truthfulness of data values)	**26**	More than x% of records have future dates	
	27	More than x% of records are in the highest or lowest y% of biologically plausibility (e.g., age, height, weight)	
	28	More than x% of patients have illogical date relationships (e.g., death date before birth date, dispense date occurs after death date)	
	29	The average number of times data was reported per "y" time frame was greater than “x”	
	30	More than x% of results for selected laboratory tests do not have the appropriate specimen source	
	31	Median patient-reported or sensor data values for selected variables are statistical or clinical outliers	
	32	The average number of principal diagnoses per encounter is above the threshold for inpatient and Emergency Department to inpatient.	
	33	The average number of patients increased or decreased by x% over pre-specified sequential time periods	
	34	The average amount of data per patient increased or decreased by x% over pre-specified sequential time periods	
	35	More than x% decrease in the number of records for ICD9 or ICD10 diagnosis or procedure codes or CPT/HCPCS procedure codes	

REFERENCES

1. US Food and Drug Administration. *Framework for FDA's Real-World Evidence Program*. US Department of Health & Human Services; December 2018.
2. 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.
3. Daniel G, McClellan M, Silcox C, Romine M, Bryan J, Frank K. *Characterizing RWD Quality and Relevancy for Regulatory Purposes*. Duke-Margolis Center for Health Policy; October 1, 2018.
4. Berger M, Daniel G, Frank K, Hernandez A, McClellan M, Okun S. *A Framework for Regulatory Use of Real-World Evidence*. Duke-Margolis Center for Health Policy; September 13, 2017.
5. US Food and Drug Administration. *Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices: Guidance for Industry and Food and Drug Administration Staff*. US Department of Health & Human Services; August 2017.
6. US Food and Drug Administration. *Standards for Data Management and Analytic Processes in the Office of Surveillance and Epidemiology (OSE)*. Manual of Policies and Procedures (MAPP 6700.2). US Department of Health & Human Services; 2008.
7. US Food and Drug Administration. *Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data: Guidance for Industry and FDA Staff*. US Department of Health & Human Services; May 2013.
8. National Patient-Centered Clinical Research Network. *PCORnet Data Checks v6*. https://pcornet.org/wp-content/uploads/2019/09/PCORnet-Data-Checks-v61.pdf; May 22, 2019.
9. Sentinel. Data Quality Review and Characterization. https://www.sentinelinitiative.org/sentinel/data-quality-review-and-characterization. Accessed September 2019.
10. US Food and Drug Administration. *21CFR820—Quality System Regulation*. US Department of Health & Human Services; April 1, 2018. Available here: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?CFRPart=820&showFR=1&subpartNode=21:8.0.1.1.12.1.
11. Callahan TJ, Bauck AE, Bertoch D, et al. A Comparison of Data Quality Assessment Checks in Six Data Sharing Networks. *EGEMS (Wash DC).* 2017;5(1):8.
12. Franklin J, Glynn R, Martin D, Schneeweiss S. Evaluating the use of nonrandomized real-world data analyses for regulatory decision making. *Clin Pharmacol Ther.* 2019;105(4):867-877.
13. Oracle. SQL Language Reference: constraint. https://docs.oracle.com/en/database/oracle/oracle-database/18/sqlrf/constraint.html#GUID-1055EA97-BA6F-4764-A15F-1024FD5B6DFE. Published 2019. Accessed July 18, 2019.
14. Kiefer C. *Assessing the Quality of Unstructured Data: An Initial Overview.* 2016.
15. US Food and Drug Administration. *Use of Electronic Health Record Data in Clinical Investigations*. Silver Spring, MD: US Department of Health & Human Services; July 2018.
16. Rubin D. Inference and missing data. *Biometrika.* 1976;63(3):581-592.
17. Patient-Centered Outcomes Research Institute. *Data Quality and Missing Data in Patient-Centered Outcomes Research Using EMR/Claims Data Meeting Summary*. December 2015.
18. Bhattacharya S, Dunham AA, Cornish MA, et al. The Measurement to Understand Reclassification of Disease of Cabarrus/Kannapolis (MURDOCK) Study Community Registry and Biorepository. *Am J Transl Res.* 2012;4(4):458-470.
19. Gender Domain and Vocabulary. Observational Health Data Sciences and Informatics. https://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:gender. Published 2016. Updated March 12, 2016. Accessed July 25, 2019.
20. Center for Effectiveness & Safety Research. http://cesr.kp.org/en/. Accessed July 2019.
21. Godlove T, Ball AW. Patient Matching within a Health Information Exchange. *Perspect Health Inf Manag.* 2015;12(Spring):1g.
22. Dusetzina S, Tyree S, Meyer A, Meyer A, Green L, Carpenter W. *Linking Data for Health Services Research: A Framework and Instructional Guide*. Rockville, MD: Agency for Healthcare Research and Quality; 2014.
23. Health Information Technology for Economic and Clinical Health (HITECH) Act, 42 U.S.C. 17902(2009).
24. IQVIA. *Physician Office Usage of Electronic Health Records Software*. May 2018.

25. Sentinel Coordinating Center. lkp_all_flags.sas7bdat. https://dev.sentinelsystem.org/projects/QA/repos/qa_package/browse/inputfiles/lkp_all_flags.sas7bdat. Updated January 26, 2019. Accessed August 15, 2019.
26. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. *Estimands and Sensitivity Analysis in Clinical Trials E9(R1)*. June 2017.
27. The Office of the National Coordinator for Health Information Technology (ONC). What are Patient-Generated Health Data? https://www.healthit.gov/topic/otherhot-topics/what-are-patient-generated-health-data. Published 2018. Updated January 19. Accessed July 25, 2019.
28. US Food and Drug Administration. FDA's MyStudies Application (App). https://www.fda.gov/drugs/science-research-drugs/fdas-mystudies-application-app. Updated January 28, 2019. Accessed 2019.
29. US Food and Drug Administration. CDER Patient-Focused Drug Development. Published 2018. Updated December 18, 2018. Accessed 2019.
30. Clinical Trials Transformation Initiative. Mobile Clinical Trials (MCT). https://www.ctti-clinicaltrials.org/programs/mobile-clinical-trials. Accessed 2019.
31. Brajovic S, Blaser DA, Zisk M, et al. Validating a Framework for Coding Patient-Reported Health Information to the Medical Dictionary for Regulatory Activities Terminology: An Evaluative Study. *JMIR Med Inform.* 2018;6(3):e42.
32. Daniel G, McClellan M, Colvin H, Silcox C, Sharma I, Bryan J (Eds). *Mobilizing mHealth Innovation for Real-World Evidence Generation. (2017)*. Duke-Margolis Center for Health

Policy. Retrieved September 24, 2019, from: https://healthpolicy.duke.edu/sites/default/files

/atoms/files/mobilizing_mhealth_innovation_for_real-world_evidence_generation.pdf; September 2017.

33. US Food and Drug Administration. Clinical Outcome Assessments (COA): Frequently Asked Questions. https://www.fda.gov/about-fda/clinical-outcome-assessments-coa-frequently-asked-questions. Published 2019. Updated February 15, 2019. Accessed September 12, 2019.
34. Active Tasks. ResearchKit. http://researchkit.org/docs/docs/ActiveTasks/ActiveTasks.html. Updated August 30, 2018. Accessed September 12, 2019.
35. McIninch J, Datta S, DasMahapatra P, et al. Remote Tracking of Walking Activity in MS Patients in a Real-World Setting (P3.209). *Neurology.* 2015;84(14 Supplement):P3.209.
36. PatientsLikeMe. In: Duke-Margolis, ed 2019.
37. Consumer Technology Association. Health Fitness & Wellness Standards. https://www.cta.tech/Research-Standards/Standards.aspx?cat=HealthFitness_Wellness. Accessed September 9, 2019.
38. Open mHealth. Schema library. http://www.openmhealth.org/documentation/#/schema-docs/schema-library. Accessed September 9, 2019.
39. Clinical Trials Transformation Initiative. *CTTI Recommendations: Advancing the Use of Mobile Technologies for Data Capture & Improved Clinical Trials*. 2018.
40. Clinical Trials Transformation Initiative. *CTTI Recommendations: Developing Novel Endpoints Generated by Mobile Technology for Use in Clinical Trials*. June 2017.
41. Consumer Technology Association. *ANSI/CTA Standard—Physical Activity Monitoring for Fitness Wearables: Step Counting*. October 2016.
42. US Food and Drug Administration. Clinical Outcome Assessment (COA) Qualification Program. https://www.fda.gov/drugs/drug-development-tool-qualification-programs/clinical-outcome-assessment-coa-qualification-program. Published 2019. Accessed July 18, 2019.
43. Coons S, Kothari S, Monz B, Burke L. The patient-reported outcome (PRO) consortium: filling measurement gaps for PRO endpoints to support labeling claims. *Clinical Pharmacology and Therapeutics.* 2011;90(5).
44. California State Legislature. *AB 375, Chau. Privacy: personal information: businesses*. The California Consumer Privacy Act of 2018; 2018.
45. The European Parliament and the Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union.* 2016;59(L119):1-88.

46. Princeton University Center for Data Analytics & Reporting. What is a Data Model? Published 2019. Accessed July 18, 2019.
47. Winglee M, Valliant R, Scheuren F. *A Case Study in Record Linkage.* Vol 31. 2005.
48. Hermansen SW. *Relational Database Schemes and SAS Software SQl Solutions.* 1999.
49. Nanyang Technological University. A Quick-Start Tutorial on Relational Database Design. https://www.ntu.edu.sg/home/ehchua/programming/sql/relational_database_design.html. Accessed September 2019.