Characterizing RWD Quality and Relevancy for Regulatory Purposes

October 1, 2018

Duke | MARGOLIS CENTER *for* Health Policy

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 University's Robert J. Margolis, MD, Center for Health Policy 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.

ACKNOWLEDGEMENTS

The Duke-Margolis Center would like to thank several individuals for their contributions to this white paper. The paper would not have been possible without the months-long collaboration of the working group members listed on page two. Their expert perspectives, open discussion, and thoughtful feedback on working drafts were indispensable and we are grateful for their support. We would also like to thank the participants in the expert workshop we held in Spring 2018 to discuss the issues around characterizing real-world data for regulatory use, as well as the advisory group members of the RWE Collaborative who provided their support and input throughout. Finally, the Center wishes to thank and acknowledge Kerry Stenke from the Duke Clinical Research Institute for her support in developing graphics for this white paper. 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 RWE Collaborative members including Eli Lilly and Company; Genentech, a member of the Roche Group; GlaxoSmithKline; Johnson & Johnson; Novartis; and Teva. For more information on the RWE Collaborative, visit healthpolicy.duke.edu/real-world-evidence-collaborative.

Mark B. McClellan, MD, PhD, is an independent board member for Johnson & Johnson 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, MITRE, and the National Institute for Health Care Management.

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

WHITE PAPER

CHARACTERIZING RWD QUALITY AND RELEVANCY FOR REGULATORY PURPOSES

AUTHORS

Gregory Daniel
Duke-Margolis Center for Health Policy

Christina Silcox
Duke-Margolis Center for Health Policy

Jonathan Bryan
Duke-Margolis Center for Health Policy

Mark McClellan
Duke-Margolis Center for Health Policy

Morgan Romine
Duke-Margolis Center for Health Policy

Katherine Frank
Duke-Margolis Center for Health Policy

WORKING GROUP

Aylin Altan
OptumLabs

Marc Berger
International Society for Pharmacoeconomics and Outcomes Research

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

Paul Bleicher
OptumLabs

William Capra
Genentech, Inc.

Kourtney Davis
GlaxoSmithKline plc

Riad Dirani
Teva Pharmaceuticals

Brande Ellis Yaist
Eli Lilly and Company

Shannon Ferrante
GlaxoSmithKline plc

Brad Hammill
Duke University

Morgan Hanger
PatientsLikeMe

Stacy Holdsworth
Eli Lilly and Company

Kristijan Kahler
Novartis Pharmaceuticals Corporation

Sally Okun
PatientsLikeMe

Michael Pencina
Duke University

Kristin Sheffield
Eli Lilly and Company

Eileen Thorley
PatientsLikeMe

Lisa Wruck
Duke University

INTRODUCTION

Real-world data (RWD) refers to data that are routinely collected and pertinent to patient health status and/or the delivery of care. Examples include electronic health records (EHRs), insurance claims data, and patient-generated health data, as well as socio-economic, environmental, genomic, and other emerging types of data. Real-world evidence (RWE) is evidence derived from RWD through the application of research methods. For regulatory applications, RWE can further be defined as clinical evidence regarding the use and potential benefits or risks of a medical product derived from analysis of RWD. This evidence can be used to characterize health status and measure a treatment's effectiveness and/or safety in real-world settings,* providing insights that traditional randomized controlled trials (RCTs) may not readily capture in a reasonable timeframe, in natural settings, or within relevant populations.

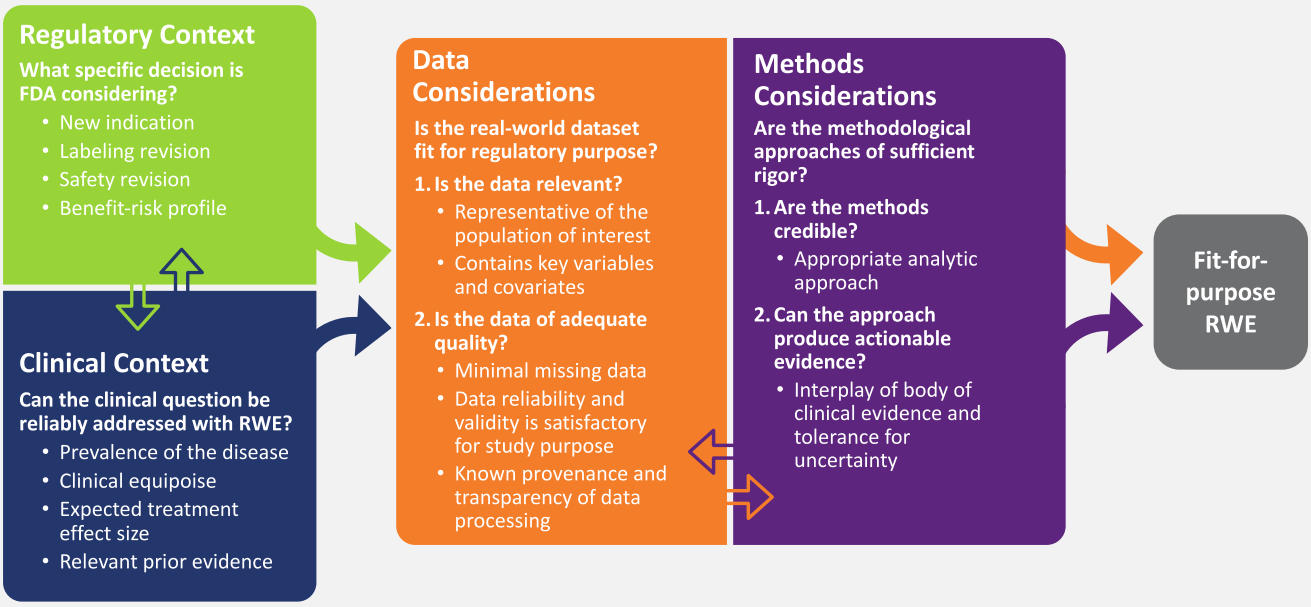
> ***Real-World Data (RWD)*** *are data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources*
>
> ***Real-World Evidence (RWE)*** *is evidence derived from RWD through the application of research methods. For regulatory applications, RWE can further be defined as clinical evidence regarding the use and potential benefits or risks of a medical product derived from analysis of RWD.*
>
> *A FRAMEWORK FOR REGULATORY USE OF REAL-WORLD EVIDENCE, September 2017*

The increasing availability of RWD and RWE has encouraged policymakers at the U.S. Food and Drug Administration (FDA) and stakeholders across the biomedical community to explore how RWD and RWE can be better integrated into drug development and regulatory review. Recent legislative efforts such as the 21st Century Cures Act[1] and the sixth Prescription Drug User Fee Act (PDUFA VI)[2] seek to answer these questions by establishing priority areas in which FDA should explore the potential use of RWE to support new indications for an approved drug and to satisfy post-approval study requirements. These regulatory use cases represent clear steps forward in the development and use of RWD, potentially allowing for more efficient and meaningful evidence generation that is better reflective of patient populations and their clinical care.

The Duke-Margolis Center for Health Policy, under a cooperative agreement with FDA and in conjunction with a group of expert stakeholders, published a framework in 2017 that addressed the considerations needed for the development of RWE that is fit for regulatory purposes.[3] The framework, an updated version of which is presented below, proposes that developing RWE that is fit for regulatory purposes should be guided by the interplay of four distinct sets of considerations: the regulatory question a sponsor is seeking to address, the clinical context within which RWE is being generated, the availability of RWD of appropriate relevancy and quality, and the application of trusted methods for turning RWD into actionable evidence (Figure 1). The data considerations box in Figure 1

* RWE studies may be produced from retrospective or prospective study designs and either with or without randomization however the data used in these studies should be real-world data routinely collected for non-research purposes. In reality, there is a continuum from clinical trial data to real-world data. For example, registries often include a hybrid of RWD and protocol-driven data, however secondary use of this data is generally still considered within the scope of RWD.

Figure 1. Considerations for generating RWE fit for regulatory purposes



As manufacturers consider a RWE development strategy to support regulatory use, there are a number of considerations that should be addressed to ensure that an RWE approach is sensible. First, it is critical to examine the intended regulatory use and the clinical context within which RWE will be developed. Second, the strength of available RWD data sources and study methods for generating RWE that is fit for regulatory purposes must be considered. Matching data sources and appropriate methods to answer specific clinical and regulatory questions will result in different "types" of RWE for different use cases (figure modified from the 2017 paper).

asks if the real-world dataset is meaningful, valid, and transparent*, and therefore appropriate to answer a specific regulatory question in a particular clinical context (i.e., fit-for-purpose). Fit-for-purpose RWD should be characterized as robust and representative of the population of interest, as well as of requisite quality to accurately capture critical endpoints and covariates.[4]

This paper expands on the 2017 framework's data considerations by further detailing the concept of fit-for-purpose RWD, a holistic assessment that includes characterizing both the relevancy and the quality of the RWD needed to produce RWE that can support a regulatory decision. These recommendations build on current FDA guidance, including the 2013 guidance "Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data" and the 2017 guidance "Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices".[5,6] The content of other frameworks that grade the quality of data are also referenced.[7, 8]

* Schneeweiss et al. (2016) described a framework for generating RWE fit for decision-making that introduced the MVET concept (meaningful, valid, expedited, and transparent). These concepts apply to RWE and encompass both data characterization and method selection, so not all of the components of MVET are within the scope of this paper. [doi:10.1002/cpt.512]

FIT-FOR-REGULATORY-PURPOSE RWD

Determining if a real-world dataset is fit-for-regulatory-purpose is a contextual exercise, with assessments of specific data characteristics contributing to an overall determination of what meaning and degree of confidence can be derived from the resulting real-world evidence. A data source that is appropriate for one purpose may not be suitable for other evaluations. For example, a large dataset that reveals critical insights about the safety profile of a new psychotropic drug may be inadequate to study potential indication expansions. While RWE offers the potential for novel and broader insight into safety and efficacy profiles of medical products, evaluations of data relevancy and quality will be critical to support claims of causal inference of treatment effect and internal validity of study results, particularly for non-randomized studies. Objective measures to assess data quality such as accuracy, validity, and completeness, as well as information on the original data collection procedures and any subsequent transformations, are critical to fully understanding the strengths and limitations of the data for the interpretation of results.

Various frameworks have been proposed both to classify data by source and to grade the quality of data along standardized dimensions.[9,10] This paper builds upon this previous work by identifying the various archetypes of RWD and suggesting technical documentation that investigators can provide on data provenance and processing for RWD. This paper will highlight documentation that is unique or particularly important to RWD as compared to traditional RCTs, addressing areas of likely concern and providing suggestions on what to document to properly report data quality and relevancy. For any specific study using RWE, some of the recommendations presented in this paper may not apply, as the appropriateness of a real-world dataset is contingent upon the specific question of interest and the existing body of evidence available.

UNDERSTANDING DATA RELEVANCY AND QUALITY

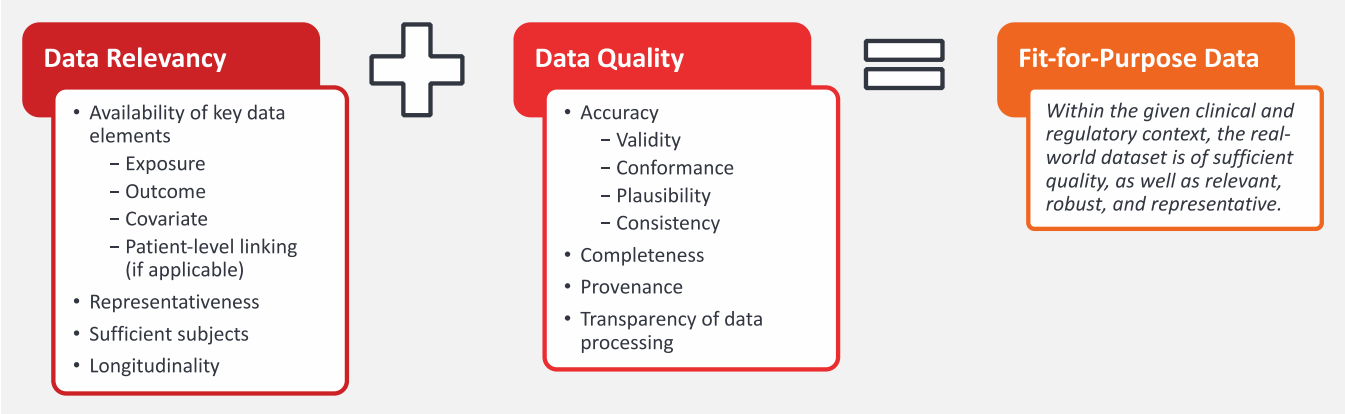
A real-world dataset should be evaluated as fit-for-purpose on dimensions of data relevancy and data quality for a potential regulatory decision within the context of a specific disease state or therapeutic area (see Figure 2).

DATA RELEVANCY DIMENSIONS

A real-world dataset is relevant if it is robust and representative of the population of interest. Data relevancy dimensions reflect whether a data set can answer the regulatory question in a clinical context of interest, outside the data quality assessment. Evaluations of these dimensions are often focused on the potential for selection bias:

- Are the patients in the dataset representative of the population of interest (i.e., patients using or who will be using the medical product)?
- Are critical data fields representing exposures, covariates, and outcomes present? If not, are these variables able to be algorithmically derived using data fields that are present?
- If more than one data source is required, are data fields present that permit accurate linking at the patient-level?
- Are there sufficient persons and follow-up time in the data source to demonstrate the expected treatment effect including adequate capture of potential safety events?

Figure 2. Data relevancy and quality are equal components of a fit-for-purpose real-world dataset



DATA QUALITY DIMENSIONS

Data quality dimensions characterize the accuracy, completeness, and transparency of RWD and seek to address potential information bias which could impact internal validity:

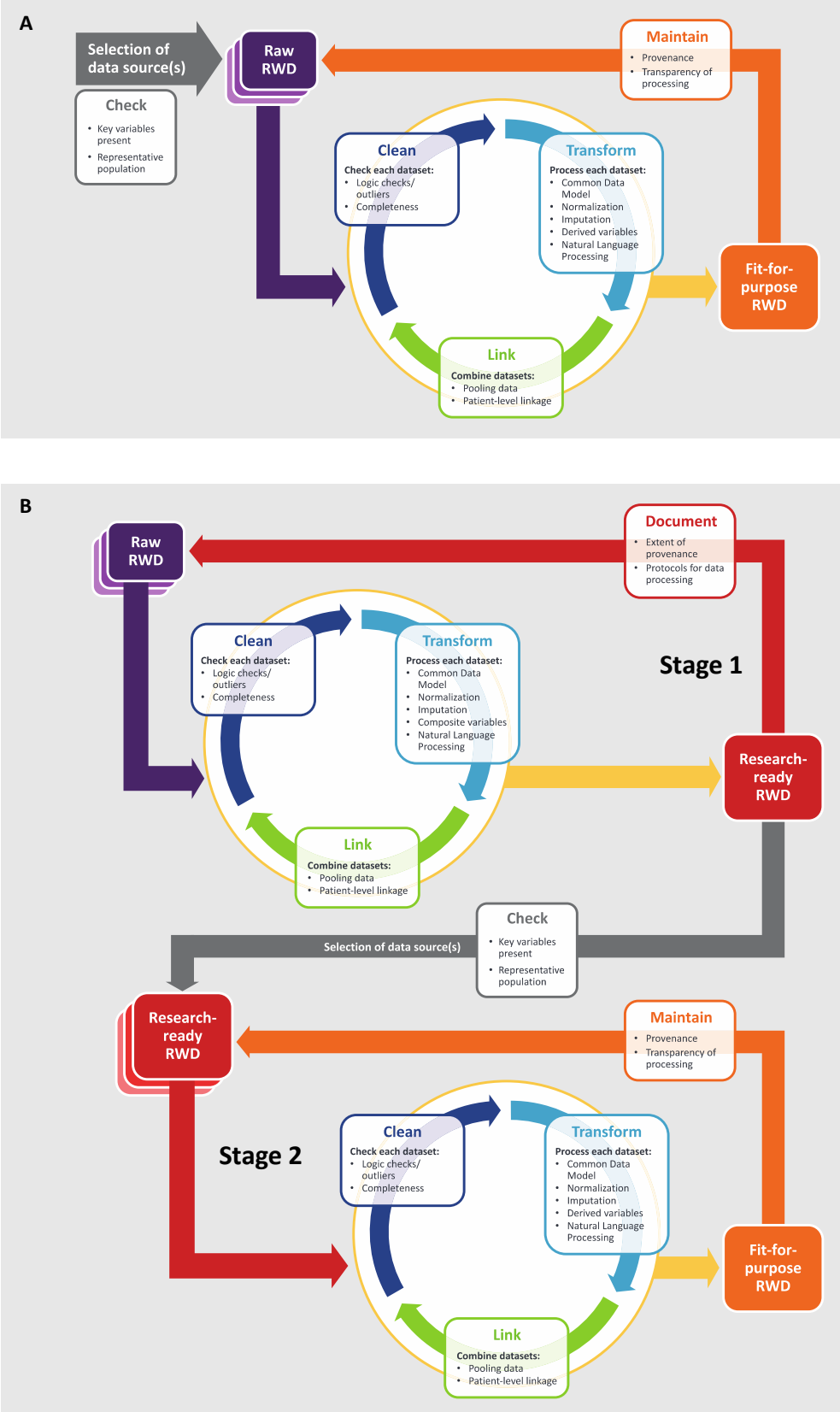
- Are the data accurate? Common measurements/surrogates of accuracy include testing the validity of the data elements and any algorithms used to transform the data, checking the logical plausibility of the data (e.g. a lab result is within the limits of biological possibility), and examining the data consistency for each patient (within related data fields and over time) as well as the conformance of the data to any applicable internal standards or external data models.
- How complete are the data? Identifying the extent and mechanism of missingness is important to understand the potential bias introduced by missing data and to identify methods to compensate. Are the data measured but not available (e.g., the patient had a laboratory test, but the result is not in the data source available for research) or was it not captured during a particular instance of routine care?
- Are the provenance and transformations performed on the data transparent as data move from point of collection into various databases? This permits key exposure, covariate, and outcome variables to be evaluated with respect to source data. For data that can be found in multiple fields, this may include specifying where in the record the data were extracted, not just in which record.

CHARACTERIZING RELEVANCY AND QUALITY THROUGHOUT THE DATASET CREATION PROCESS

PROCESSING RAW RWD INTO FIT-FOR-REGULATORY-PURPOSE RWD

The process of producing a fit-for-purpose real-world dataset begins with the selection of one or more data source(s). More typical raw real-world data sources include electronic health records (including structured data, free text, images, lab results, prescriptions, etc.), administrative medical and pharmacy claims data, data from medical devices, and consumer data (data from wearables and other consumer devices, websites, social media, pharmacy loyalty cards, etc.). While not the focus of this paper, data from emerging sources such as bioinformatics and genomics databases, from public health databases such as data from environmental sensors, and from other evolving technologies will impact our understanding of background risk, exposures, health outcomes, and may be used as sources of RWD for regulatory decision-making. Each source and type of RWD has different benefits and limitations in the dimensions of data quality and relevancy (see Appendices A and B for a listing of the types of data found in common examples of real-world data sources, as well as a discussion of the benefits and challenges in using data from those sources).

Figure 3. The process of making a fit-for-purpose real-world dataset



When selecting one or more data sources, an investigator should show that the subset of data that they will use provides the relevant outcomes, exposures, and sufficient covariates to address specific regulatory questions at the required level of certainty. In some cases, study investigators may choose to start with raw RWD collected directly from provider systems, payers, etc. All of the processing steps will then be done by the study investigator and documented with a clear data management plan (DMP) with versioning control (see Figure 3a).

In other cases, the raw RWD may be sourced, aggregated, and transformed by third-party entities to create so-called research-ready RWD (see Figure 3b), which can then be used by study investigators to produce fit-for-purpose RWD for their particular study. In this case, it is likely that the data will go through a number of iterative steps aimed at producing a fit-for-purpose real-world dataset.

Within these steps, raw RWD are cycled through the stages of data cleaning, linking, and transforming as needed. For example, study investigators utilizing patient-level linked EHR and personal digital health application[*] data may require cleaning and transforming the data fields relevant for exposure and patient linking first and then additional cleaning and transforming on the outcome and covariate data in the linked dataset. Standard operating procedure (SOP) documentation should be maintained by data aggregators to describe the cleaning, transforming and linking of data into a research-ready database, along with summaries of data accuracy and completeness. Curation of study-specific analytic datasets from these research-ready databases may then require additional data processing by the study investigators to answer the question of interest, such as combining data fields to form composite endpoints, linking additional sources of data to gain clinical granularity or longer-term outcomes, or additional study-specific data cleaning.

For processing performed by the study investigators, it is important that the investigators maintain documentation regarding data provenance and processing (including detailed DMPs) to be made available for auditing, similar to the data management of RCTs. Transparency and documentation of these steps and any decisions and assumptions regarding the necessary cleaning, transforming, and linking of data are critical to characterizing the final fit-for-purpose dataset used by sponsors. The following section describes the steps in processing RWD into a fit-for-purpose dataset more fully and discusses the recommended documentation for characterizing the RWD relevancy and quality.

SELECTION OF DATA SOURCE(S) AND INITIAL EVALUATION OF DATA QUALITY

Investigators should provide a rationale for selecting data source(s) for a specific use case and regulatory decision. Statistical analysis plans and study protocols should clearly demonstrate RWD is fit-for-purpose by pre-specifying anticipated concerns with the RWD (e.g. known or suspected selection and information bias) to discourage post hoc analysis.[11] Background information on the clinical context of the disease and treatment (e.g., method(s) of diagnosis, preferred and actual treatment patterns for the disease(s) of interest) and the degree to which such information is collected in the proposed data sources may be helpful in explaining the rationale for data source selection and potential bias concerns.

RECOMMENDED DOCUMENTATION

- Confirmation that the RWD contains the pre-identified critical data fields as well as a sufficient and representative population for generalization of results to the population of interest
- The extent of traceability and provenance of the data from initial collections to when the investigators acquired it.
- Initial assessment/discussion of potential selection and information bias associated with the selected data source

The selection process should include an evaluation of potential systematic bias that is consequential to the

[*] A detailed description of personal digital health applications can be found in Appendix B.

analysis. Examples of this include practice variations that influence the documentation of relevant comorbidities, latent factors that affect patient motivation to actively track and contribute data[12], measurement validity of prescriptions written and observed in EHRs, influence of formulary rules and health plan coverage policies, EHR macros that autofill outdated information about patients and medical encounters, diagnoses that are commonly upcoded[13] and/or pharmacy dispensing data that indicate exposure to treatment.[14]

Next, the accuracy, recency, and completeness of record for critical variables and endpoints should be carefully considered and documented. Guidelines and checklists for selection of data sources (such as the ISPE Guidelines for Good Pharmacoepidemiology Practice (GPP)[15] and the Good ReseArch for Comparative Effectiveness (GRACE) principles)[16] emphasize sufficient completeness, detail, and validity of exposure and primary outcome variables, as well as the presence of key confounders and effect modifiers. The documented rationale for the selection of RWD sources should reflect this analysis. Pre-specified time intervals for data collection should be made clear as well as a description of the variables and assessments that can confirm the sufficiency of the size, granularity, and representativeness of data collected on patients, exposures, and outcomes.

CLEANING, TRANSFORMING, AND LINKING RWD

Once real-world data sources are selected for use in a real-world study, the process of cleaning, transforming, and linking can begin. DMPs should be used to document how study data were collected and stored, which personnel

CHARACTERIZING DATA SOURCES AS "RESEARCH-READY"

Data originators, aggregators, research organizations, and others may start the cleaning, transforming, and linking process as they collect data into "research-ready" databases and registries that will then be used by sponsors and researchers (see Stage 1 in Figure 3b). While the sponsor has the final responsibility to collect all needed information to characterize the dataset as fit-for-purpose, it will facilitate the process if these initial efforts are well documented.

These database owners can increase the credibility and value of their data by making available the characteristics of the population represented in the database, as well as information on the availability of specific data elements and individual assessments of quality of those elements. Information on audit processes, accuracy loss that may occur during standardization, algorithmic transformations, and workflows during original collection that may introduce bias are critical for sufficiently characterizing the appropriateness of particular data sources to make a fit-for-purpose dataset. Records regarding the traceability for key data points of the study are important to assessing the quality of the data source(s).

Pre-Certifying a Research-Ready Database

Since sponsors and researchers may often draw their specific RWD datasets from continually updated and curated research-ready real world databases, it may be more efficient to establish a process by which those databases become "pre-certified" for regulatory applications. A pre-certification process would confirm that the database follows appropriate SOPs and quality controls, and that it has the proven capability to produce real-world datasets fit-for-regulatory purpose. This would take developing consensus on the key features of the pre-certified database, the process for becoming 'pre-certified', and the most appropriate certifying body. Such a process could simplify the reporting and documentation needed for Stage 1.

are responsible for maintaining data integrity, and standard operating procedures for working with the data. DMPs may be revised as needed to reflect changes to study procedures, and a version history should be maintained.

Investigators should include source documentation for individual data elements and SOPs for data collection, transformations, and audits within the DMP.[17] Actual data verification procedures may depend on the regulatory context and question of interest. Documentation should include the original sources that collected the data and the method of data capture (e.g., clinical data abstraction, EHR integration, linkage to claims data, and personal digital health application data).[18] The frequency and type of any data error corrections or changes in data adjudication policies implemented by the data holders during the relevant period of data collection should be noted.

Cleaning RWD

Data cleaning prepares RWD for analysis by identifying errors in the data against a known standard and removing or altering the data to conform to that standard. The need for RWD to be cleaned using logic checks, assessments of completeness, duplicates, and evaluation of data collection errors (including entry, measurement, integration, and summarization errors) are well documented.[19]

RECOMMENDED DOCUMENTATION

- Documentation of the cleaning process, including validation of data against transparent standards and removal of erroneous data
- Summary measures of data completeness and identified errors

The Reporting of Studies Conducted using Observational Routinely-collected Health Data (RECORD) statement (an addition to the original Strengthening the Reporting of Observational Studies in Epidemiology guideline) advocates for clear disclosure of all operations performed on the data.[20] This is done to ensure repeatability of the study design and reproducibility of findings. Study investigators should be clear about their strategies to minimize missing data and assumptions regarding why data may be missing. In the context of patient experience data, FDA recommends providing summary statistics of missing data frequencies and percentages, possibly including stratification by important subgroups.[21]

Transforming RWD

Data transformations include converting a dataset into a common data model, de-identifying data, normalizing recorded values, classifying clinical events, imputing missing data, and using algorithms to calculate composite or summary variables. Further data transformations may include summarizing free text using natural language processing and converting raw digital signals to numerical summaries. After these transformations, it is critical to analyze and document that the data are still appropriately accurate (including any potential bias introduced by the transformations[*]), representative, and sufficient.

RECOMMENDED DOCUMENTATION

- Transformation procedures for RWD should be documented, including the purpose, historical uses, and any performance metrics
- Critical transformations such as data imputation, algorithmic data summarization, and de-identification may require more information on the changes to the data post-hoc

It is important that clear documentation and versioning information are provided when data transformation algorithms and common data models are used so that reviewers can easily trace these changes and understand

[*] Transformations can induce more recognizable patterns in values to make analysis easier, however, it can also reduce granularity.

their potential impacts when interpreting heterogeneity of records across data sources. Changes that can affect data over place and time, such as changes to prescribing practices, formularies, coding systems, and diagnostic guidelines, may require domain knowledge and ad-hoc approaches to ready data for analysis.[22] Detailed notes on the decisions made and the methodologies used will be needed for external reviewers to understand how these transformations may affect the relevancy and quality of the data.

Decisions on the methods used to impute missing data should be made transparent, including assumptions regarding the underlying mechanism of missing data (e.g., missing at random versus missing not at random).[23]

Often investigators will use algorithms to approximate RWD outcome measures and important predictor variables that are not captured directly or with enough granularity. Documentation of prior algorithm validation studies, either previously published or validated via an embedded study with a subset of the real-world dataset being used, should be provided or cited to characterize the accuracy of the transformed data.

All real-world studies should follow the appropriate governing authorities for patient privacy (including the Health Insurance Portability and Accountability Act (HIPAA) and the Common Rule, if applicable) and document any effects on data quality or relevancy. This is of particular importance when using de-identified data, which is created through a transformation process that requires a precise balance between protecting patient privacy and maintaining data granularity and richness. Appendix B includes more detailed information on how the de-identification process may affect the characterization of fit-for-purpose RWD.

Linking RWD

There are different purposes for combining data from different sources in order to make a fit-for-purpose dataset. The type of documentation will therefore differ depending on the purpose and methods for linking data sources.

RECOMMENDED DOCUMENTATION

- Data linkages constitute either pooling common datasets to increase sample size or patient-level linking of disparate datasets to increase data richness
- Performance metrics for procedures that link datasets should be reported
- Critical differences in each distinct dataset should be reported, including varying methods of measurement for common data fields, selection bias, and changes in standards
- Procedures for adjudicating conflicting data for unique individuals or observations should be reported

For example, data pooling combines datasets that contain similar data fields among unique patients to increase sample size. Pooling data in order to extend the sample size (and potentially the representativeness) of the study is generally straightforward, but must account for the variation and biases that each distinct data set contributes based upon how it was recorded and pre-processed. These may include coding and diagnostic discrepancies across different provider and health payer systems. Alternately, sponsors may decide to use a distributed database design that keeps the data in separate files maintained by individual data providers; in this model each provider utilizes the same analysis plan and the analytical results from each are pooled by the central sponsor or researcher. A prominent example of the distributed data network is the Sentinel System where analytical queries are sent to collaborating insurers and health care systems that map their RWD to a common data model.[24] In these types of systems, it is the results, not the data, that is pooled.

Alternatively, investigators may link disparate data sources at the patient-level to improve the richness of the data available for analysis. In this case, sponsors must document quality characteristics of the individual data sources, the methods and performance metrics of the linkage procedure, and any loss of data quality or relevancy due to the linking process. Sponsors must also have a transparent rationale for reconciling differences when more than

one data source reports the same data. For linkages that increase data richness, transparency of the linking process is required to assess the nature of potential selection bias. Some data linkages may systematically drop out patients with certain characteristics who are missing key information across the selected data sources.

Either data linking approach may encounter discordant data entries for the same individual (e.g. an individual's basal metabolic rate was recorded differently in the same time period across two different RWD sources). In such cases, investigators must clearly document their SOP for reconciling and validating discordant data entries for unique patients.

CHARACTERIZING THE FINAL REAL-WORLD DATASET AS FIT-FOR-PURPOSE

In addition to the detailed DMP discussed above that discloses data provenance, data processing procedures, and all critical assumptions made regarding collection and preparation of the RWD, a final characterization of fit-for-purpose RWD should include numerical and qualitative summaries of data relevancy and quality.

While the clinical and regulatory context will change specific reporting requirements, the core theme is that each RWD-supported submission should have enough documentation to transparently characterize the relevancy and quality dimensions of the real-world dataset to the specific regulatory decision at hand. This should include RWD-specific documentation such as the following:

- Historical use and prior data management documentation of RWD sources;
- Assessments of selection bias from data sources;
- Assessments of information bias from data sources;
- Impact of assumptions and procedures from data cleaning, transformation, de-identification, and linkages;
- Assessment of changes in key data element capture and coding over time;
- Measurements of accuracy for critical data fields, such as consistency with source, sensitivity, and specificity of calculation and/or abstraction;
- Historical or verified validity measures of critical data fields; and
- Assessments of data completeness by field and over time.

Note that these details refer only to RWD-specific concerns on characterizing datasets for regulatory decision-making. Previous work by the FDA and others should be referred to for more general guidance on data submissions.[25,26]

CONCLUSION

RWD can open novel avenues of insight into patient health while informing regulation, improving therapeutic use, and illuminating care pathways. The FDA has taken great strides to harness such data for postmarket safety surveillance, and will continue to explore its potential use for regulatory decision centered on questions of effectiveness. This will require a clear characterization of how and when RWE is fit for a variety of regulatory uses beyond safety.

Through collaborative work with expert stakeholders, we propose that this characterization is split into two sets of dimensions, one set regarding relevancy of the data to the specific regulatory question at hand and the other on the quality of the selected dataset. Moving forward it will be critical for study sponsors to document data relevancy and quality characteristics throughout the process of generating, collecting, and processing such data for analysis and regulatory decision-making. This will require study sponsors and third-party aggregators to work toward robust and reproducible RWD-generating procedures with transparent data management capabilities. Future work is needed to identify best practices for responsible curation of fit-for-purpose RWD.

APPENDIX A: EXAMPLES OF COMMON REAL-WORLD DATA SOURCES

Table 1. Examples of common real-world data sources

Data Source	Data Characteristics	Examples of Included Data Types
Electronic Health Record (EHR)	A combination of structured (although not standardized) and unstructured data fields that contain data from clinical encounters.	• Diagnoses, symptoms, and treatments • Diagnostic test results (imaging, genetic tests, medical device data, etc.) • Ordered/written prescriptions • Demographics • Patient experience data • Clinical narratives
Administrative Claims and Enrollment	Structured data fields from claims submitted by health care providers—e.g., individual physicians, inpatient facilities, outpatient facilities, or other service providers—for reimbursement of covered services. Note that many of the large commercial payers have additional clinical data such as lab results from national laboratories who provide services for their members.	• Diagnosis and procedure codes • Dates of service for all encounters; inpatient lengths of stay • Outpatient pharmacy dispensing data • Provider and facility information • Plan/benefit information, including enrollment windows • Demographics
Personal Digital Health Applications	A combination of structured (although not standardized) and free text fields and sensor recordings that contain data reported directly by the patient into websites or apps or generated by medical and consumer devices.	• Patient experience data • Data from personal medical devices • Sensor data from consumer devices • Socioeconomic and demographic data • Over-the-counter medicine lists • Meal tracking • Activity levels
Public health databases	Generally consists of structured datasets made available for analysis, with various degrees of access, by governments, non-profits, and the research community that contain data not found in EHRs or administrative claims but do affect the public health.	• National Death Index • EPA Air Quality Systems (AQS) Data Mart • HRSA Area Health Resource File (AHRF) • AHRQ Healthcare Cost and Utilization Project
Emerging sources	Data sources that thematically capture data and information about patient physiology, biology, health, behavior, or their environment that have not been substantially assessed or validated in their ability to produce reliable and credible RWE.	• Genomics • Metabolomics • Proteomics

APPENDIX B: CONSIDERATIONS FOR ASSESSING DATA RELEVANCY AND DATA QUALITY BY DATA SOURCE

As discussed in the main text, RWD includes, but is not limited to, data in electronic health records (including structured data, free text, images, lab results, prescriptions, etc.), administrative and claims data, and data from medical devices, as well as consumer data (data from wearables and other consumer devices, websites, social media, pharmacy loyalty cards, etc.). Appendix A lays out examples of data types that can be found in common real-world data sources. Each source and type of RWD has different benefits and challenges in the data quality and relevancy dimensions. Below, we discuss some of the data relevancy and quality benefits as well as common challenges that should be addressed when documenting how a study may address potential sources of information and selection bias.

ELECTRONIC HEALTH RECORDS (EHRS)

The EHR* has become a routine tool for clinicians to collect, aggregate, and retrieve patient information, however its use as a real-world data source for regulatory decision-making is still in early stages. Approximately 70 percent of all U.S. physicians have adopted an EHR system for routine care.[27] EHRs can be rich sources of patient data depending on the clinical context and motivation underpinning EHR data entry (e.g., use for billing, patient health outcomes, physician quality assessment). It is important to be specific when referencing EHR data as it can include multiple data types, such as structured data, free text descriptive information, laboratory and imaging results, and more.

EHRs and Data Relevancy

The amount of clinical data in EHRs has greatly expanded the availability of key data fields that are available electronically as RWD, including more granular information around symptoms, risk factors, diagnosis, and treatment. This allows the study of more tightly defined exposures and more detailed clinical outcomes. Important covariates that are not available in claims data have allowed for more precise risk adjustment using EHRs as well.[28] However, data field availability is not necessarily consistent among EHR vendors or between provider systems. For example, in some systems, lab test results may be available to health care providers, but not integrated into the EHR. If key data are not present in the EHR, patient-level linking to a data source that does contain the target data may be required.

A fit-for-purpose real-world dataset must be representative of the population of interest as defined by transparent and reasonable criteria. EHR data must be examined for potential systematic bias that could affect the representativeness of the dataset. For example, if only one or two provider systems' data are being used, the patient mix may not be demographically or socioeconomically reflective of the likely users of the medical product. Physician preference and utilization patterns specific to those systems may also lead to selection bias concerns. For example, information in EHRs is often systematically biased towards diagnostics and procedures needed for billing[29] including both "downcoding" (only reporting the information required to support a specific in-patient claim code) and "upcoding" (reporting a more complex diagnosis than is supported by the evidence). Furthermore, a specialty center population may not be generalizable for certain questions, or a community hospital may not have a sufficient number of patients to answer research questions for certain conditions. Similarly, a single EHR

* FDA defines an electronic health record (EHR) system as "an electronic platform that contains individual electronic health records for patients. EHR systems are generally maintained by health care providers, health care organizations, and health care institutions and are used to deliver care. EHR systems can be used to integrate real-time electronic health care information from medical devices and multiple health care providers involved in the care of patients." An EHR is an individual patient record contained within an EHR system.

data set may be too small for rigorous analysis after patient data are filtered through inclusion and exclusion criteria. Linking data from multiple provider systems or conducting a distributed study pooling data from provider systems with a standard protocol may be a solution if there are representativeness and/or sufficiency concerns.

EHRs and Data Quality

The primary purpose of EHR data is clinical care and documentation for payments, which can indirectly create challenges for secondary research purposes.[30] Several factors can increase or decrease the accuracy of individual data elements. Information bias can be a significant concern for the validity of EHR data. Factors that can influence how physicians enter data include accepted practice patterns, EHR user interfaces, cut-and-paste, auto-filling, time proximity to use of that data in clinical decisions, decision support tools, and formulary status. For example, measurements like blood pressure may be taken more precisely in the cardiology clinic or immediately before surgery compared to the primary care setting. Data that are critical for care or can be audited as part of payment or quality checks, such as the CMS Hospital Inpatient Quality Reporting Program, are thought to be more valid and more likely to be entered carefully and/or corrected (which can affect plausibility).[31] Many EHR systems have standardized basic data fields (although this is less common for practice specialty EHR systems) and are moving towards interoperability to ease the transfer of electronic health information, but conformance to data standards can vary widely across provider systems (often because providers are not aware of them), a critical concern if pooling data. Conformance and consistency are often concerns even within a single system due to significant variation between EHRs coding systems (e.g., outpatient, hospital, labs) and/or options for where to put the data (e.g., structured vs free text). Many integrated systems who commonly use EHR data for research internally and externally as part of their mission have worked to address many of these concerns for a limited set of core data elements.

Another key quality concern with EHRs is that much of the data are contained in unstructured data fields and most current EHR systems do not systematically encode free text. Data from unstructured fields can be accessed via chart review by trained chart abstractors (e.g., nurses, tumor registrars). In addition, natural language processing algorithms are increasingly adept at converting free text to structured data fields. However, abstraction and extraction may introduce unique bias and variation into the final data set. Like any other transformation, the processing steps and assessments of these transformations will need to be documented to fully understand the accuracy and provenance of these data elements.

EHR completeness is highly variable, depending on the incentives for recording and linking, the utility of the data to the health care provider, and the user interface. Missingness of key variables in a final dataset can be high for many reasons, including incomplete linkages between specialties and across practices with a provider system, inconsistent provider data entry, and differences in how EHR systems pull in data from medical devices and

INTEGRATING MULTIPLE TYPES OF DATA

Linking different types of data sources can at times overcome the limitations of the individual types. For examples, combining EHR data with claims at the patient level allows investigators to combine a known denominator from the claims data enrollment information with the clinical richness of the EHR data. This allows more specificity in when defining the study population (e.g., confirming the indication when there is more than one for licensed a therapy) and further detail on the outcomes (e.g., lab results, severity scores, etc.)

The principal challenge is often the reduction in subject cohort size and matching patients across time and place. Sample size is especially important in order to capture enough data on physician-level variation and patient case-mix to provide statistical power for answering a question of interest, especially when treatment effects are small.*

external laboratories. Key data fields such as mortality often only include deaths that occurred at the hospital. Further, it is not unusual for expected data fields such as height and weight not to be consistently recorded in an EHR even when measured in the clinic. Because EHRs will have missing data, analytic techniques, including subgroup analyses of complete case cohorts, may need to be used to estimate the impact of missing data on the question of interest. For data fields where completeness is not random, such as mortality, linking to other sources of data may be required.

It is also common for people to seek care from more than one provider and more than one health care setting. Even when patients stay within the same provider system, data may not be captured across clinical subunits. For example, inpatient care information may be missing from ambulatory and specialty EHR data, or specialty centers within hospitals which maintain specialized EHR systems may not fully interoperate with the main EHR system. Combining EHR data with claims may help identify the extent of this "leakage" and potentially fill in some gaps (see exhibit at right).

Like all real-world datasets, documentation regarding relevancy and quality of the data is critical to demonstrating that a dataset is fit-for-purpose. Particularly crucial is traceability, of both the provenance of the data (including the type of institution, EHR platform, and internal understanding of workflow and incentives that may affect coding practices) and any transformations that have occurred (including the entity who performed the transformation, what transformation was done, and at what time).[32] Many variables in an EHR dataset have multiple possible sources, and those sources may or may not agree; which data source is chosen under which circumstances and adjudication between sources is important to consider.

ADMINISTRATIVE AND CLAIMS DATA

Administrative and claims data are the most common RWD used for real-world studies to date. Because of this, many of the benefits and challenges are well understood and previously documented.[33,34]

Claims and Data Relevancy

Compared to other data source types, claims data are more likely to contain structured and standardized clinical event coding (e.g., diagnostics, Current Procedural Terminology (CPT) codes, and pharmacy dispensing) designed to support billing claims. However, since claims data are designed for billing purposes, they do not always distinguish claims intended to rule out a condition from those which confirm that condition. Claims data often lack key data fields around outcomes such as clinical details and results (e.g., vitals, lab results, key details on staging or severity of disease), requiring proxy measures (e.g. procedure codes, combinations of health care utilization plus treatment dispensings).[*] Complete mortality information is generally unavailable with the exception of Medicare claims data. Claims data also may lack necessary covariates for observational analysis, as what is coded is only what is needed for billing and reimbursement for the provided health care service. As such, diagnosis codes may be missing, incomplete, or lack concordance with concurrently collected clinical data.[35] For example, one study found that the sensitivity of ICD diagnosis codes for metastatic cancer is only about 50-70%.[36]

Depending on size, payers can be geographically diverse and offer large sample sizes, which is helpful for sufficiency but may not have the necessary follow-up if patients drop out of their plan or pass away (death is not a billable event). Also, representativeness can sometimes be a concern, depending on the size and reach of the particular payer. For example, private claims often come from employer-based insurance that covers younger and healthier employed populations and their dependents, while Medicare data is representative of Americans age 65 or older or those with specific disabling conditions. Depending on how well the payer population is

[*] There are payers that have these clinical details due to partnerships with other entities that have these data, such as Optum and Healthcore.

generalizable to the population of interest, a representative dataset may require including data from multiple payers.

Claims data has the advantage of longitudinal information about a patient's encounters with multiple provider systems and settings of care as well as pharmacy dispensing information and enrollment start and stop dates. However, there is often a significant lag (3 months or more) between when an event is recorded and the availability of complete claims data, particularly for CMS claims data, making time-sensitive analysis more difficult. In addition, changes in enrollment remain a concern. A health plan will only have information about a patient as long as they are covered. Depending on the population and the type of plan, the average length of time varies, with some patient populations cycling within the course of 6 months to 2 years.[37] The impact of the loss of information depends on the likelihood that health insurance or employment changes selectively include patients who are likely to benefit or be harmed more than people who do not switch coverage or employers. The Medicaid population is particularly transient due to tight eligibility requirements. In addition, claims data will not reflect uncovered services.

Claims and Data Quality

A great deal of methodological research has been done to validate claims data with algorithms used to accurately attribute health conditions, exposures, and outcomes for regulatory decision-making, such as the Sentinel Health Outcomes Inventory/Validations.[38] However, when combining data from different plans, care must be taken to examine how certain benefit designs may influence the prevalence of certain interventions (particularly pharmacotherapy and mental health services) and outcomes. Insurance designs such as population health management programs may also introduce unexpected confounders (e.g., programs meant to encourage systematic changes in patient behavior). In particular, it is important to keep in mind that care pathways are often largely influenced by pharmacy benefit programs and that information about what drugs are covered at preferential rates is rarely available due to confidentiality concerns.

Conformance to controlled terminology is generally high, with data in standardized ICD 9/10 and HCPCS/CPT coding, although coding/billing practices may vary by provider and payer.

For reimbursable health care services/events, completeness is typically high, however, it can differ among providers based on what payments are being requested. For example, the presence of a diagnosis code can depend on a treatment service being offered by a particular provider and whether the code is beneficial for billing. Similarly, insurance companies' ability to track provenance of patient data depends on the provider and the payer's ability to access patient clinical information. Payers can also conduct their own proprietary transformations of claims, augmenting provider data. This makes data transformation transparency key to explain potential relevancy and data quality issues.

PERSONAL DIGITAL HEALTH APPLICATIONS (PDHA) AND PATIENT-GENERATED HEALTH DATA (PGHD)

Consumer-facing personal digital health applications (PDHAs) collect a broad range of patient-generated health data (PGHD). This data is typically stored by the patient, application owner or third-party aggregator that allows the patient to view and often share this data. Platforms for PDHA typically include smartphone applications, websites, consumer wearables, and personal health records. PGHD can also be found in EHR and administrative data sources, such as from the results of a patient questionnaire or data collected from a medical device. The collection method, format, and use of data stored in PDHA can vary widely from the way PGHD is used in EHR and administrative data sources. Evaluations of real-world studies that include PGHD therefore must take into account the unique features of the data source as well as the particular data types.

PGHD represents a range of patient experience data: patient-reported outcomes (PROs); data from medical devices; data from consumer-grade wearables, websites, social media; pharmacy loyalty cards; etc. Efforts are

underway to characterize and validate PGHD, including the FDA Clinical Outcome Assessment Qualification Program, [39] Consumer Technology Associations' Health, Fitness and Wellness Subcommittee,[40] and the Critical Path Institute's Electronic Patient-Reported Outcome Consortium.[41] PGHD allows researchers the opportunity to study novel outcomes not well-suited to direct clinical measurement or not captured in clinical settings. While the use of patient experience and medical device data has been detailed in clinical literature, the data quality and relevancy of other types of PGHD derived from PDHAs can be difficult to characterize due to a lack of experience and standardized tools for assessment for a continuously growing ecosystem of data types.[42] Nonetheless, it is important to acknowledge that such data may be more accurate than data that depends on recall or testing in controlled situations. For example, a wearable that collects walking activity data continuously over a 24-hour period maybe be more meaningful than a discretely assessed 6-minute walk test administered in a clinic. It may be worth considering using newer and validated forms of PGHD as complementary data to a larger body of evidence.

PDHA and Data Relevancy

The potential to supplement clinical data with real-time biometric data and/or patient experience data promises a future of medical research that is more personalized and meaningful to patients.[43] Data from PDHA can provide novel data on exposures, outcomes, and covariates, such as a more a complete picture of the background risk of the patient for an event of interest, more relevant endpoints (e.g., steps walked, hours slept), and measurements of real-time exposure (e.g., a sensor on an inhaler recording the dose, geographic location and the time and date of use). However, PDHA may present unique logistical, legal and regulatory challenges, such as customized PDHA data standards, increased probability of re-identification, and nascent regulations for mobile health devices and applications.[44] Because consumer companies developing PDHA are generally not covered by HIPAA, careful consideration regarding how the data were acquired (e.g., through patient-mediated sharing) and the level of consent should inform the decision to use the data in a potential regulatory submission.

Longitudinality can be affected by the mode of data collection, which often requires sustained engagement from the patient. The extent of the required engagement spans from passively collected data from implanted pacemakers (which only require periodic electronic interrogation during a clinical exam), to charging, syncing, and wearing activity monitors, to answering a daily questionnaire. Patient motivation and familiarity with technology may influence drop-off rates as well as the representativeness of the dataset. Patients that demonstrate sustained collection of longitudinal data may differ significantly from those that stopped data collection early, reducing the generalizability of the final dataset to the population of interest. FDA encourages that submissions with patient experience data disclose the specific sampling methods used to collect the data, including probability and non-probability sampling, to give information about the generalizability of results.[45]

PDHA and Data Quality

Because many PDHA are novel and few data standards exist, auditing these platforms for data accuracy and conformance can be challenging. PDHA may contain high-quality PGHD including Clinical Outcomes Assessments (COAs)[46] from validated PROs and data from FDA-approved medical devices. These often have well-documented studies on the validity and precision of their measurements. FDA has demonstrated an interest in exploring, evaluating and analyzing the utility of PGHD for regulatory purposes through research collaboration agreements to explore and characterize emerging types of PGHD, including joint efforts with the Clinical Trials Transformation Initiative and PatientsLikeMe.[47] The Consumer Technology Association has started to produce voluntary standards around validity and conformance for physical activity and sleep monitors, but much of the data in consumer devices remains unstandardized and there is little transparency around the algorithms used to convert raw sensor data into the information provided to the user due to intellectual property barriers.[48] Bring-your-own-device

(BYOD) solutions can, therefore, introduce conformance and consistency concerns when patients use their own device to record important data on their physical and subjective states.[49]

Lack of widely adopted standards is a particular challenge for patient experience data outside of validated PROs. Patient registries and other organizations such as PatientsLikeMe have developed systems to capture thousands of patient experiences with various diseases, including efforts to mine PGHD and unstructured clinical trial adverse event data for concordance.[50,51] However, organizations and solutions structuring patients' data for research remain the exception. PDHAs that lack standardized collection procedures and validity assessments of their data impede characterizations of the distribution of data errors. Data extracted from PDHAs should include documentation of any data transformations to trace the provenance from raw to summarized data fields. Completeness may also be an issue when patients fail to consistently wear a device, charge the device, or miss opportunities to record their data.

As PDHAs evolve, their applicability to regulatory decision-making in the near term will likely rely on sufficient one-off validation assessments that are compelling to a particular regulatory question under consideration. The lack of standardization in PDHAs means that documentation around provenance of the data and methodological metadata about the initial data collection is particularly critical. Transparency around algorithmic transformations is also very important, though concerns around intellectual property can make this a challenge. In the absence of algorithmic transparency, validation data that clearly demonstrates an accurate and reliable measurement of the outcome of interest over time may be acceptable.

PUBLIC HEALTH DATABASES

Federal and state governments, research institutions, and others often maintain sources of real-world data for population health. For example, the National Death Index, maintained by the US National Center for Health Statistics, is a comprehensive database of dates and causes of mortality for US citizens, data that is often unavailable from traditional RWD sources.[52] Other databases capture environmental determinants of health, such as geographic differences in air quality over time. The quality of the RWD provided by these databases will depend on the origin of the data and format. Established public data sources should have extensive documentation regarding their collection methods and maintenance as well as validation studies.

EMERGING SOURCES

The sources of RWD continue to expand and will continue to pose new possibilities and challenges for stakeholder use. Data from genomics, microbiomics, metabolomics, and proteomics may offer increasingly more precise health information. Other emerging data may give us deeper insight into human behavior and our environments through the use of networked devices (Internet-of-Things) and autonomous sensors. Some proof-of-concept work is already ongoing: state-of-the-art applications have demonstrated the appeal of linking clinical and genetic data to investigate novel target identification and comparative effectiveness in oncology.[53] Air quality and socioeconomic data have been used to explain fluctuations in asthma-related emergency department visits in New York City.[54]

A thorough evaluation of the relevancy and quality of these data has not yet taken place. Additional effort will need to explore ways to maintain patient privacy when precision medicine and environmental data is linked to clinical, administrative, and patient-generate health data. For many instances of omics data, other work will be needed to establish common data quality standards, stable reference data, or diagnostic interpretations.[55] Developing such assets into a testable framework will allow more confidence in the use of these emerging sources of data.

APPENDIX C: DATA DE-IDENTIFICATION

Researchers routinely transform RWD to de-identify datasets, which allows research with personal health information (PHI) outside of the HIPAA regulations that require patient consent. While other federal, state, and institutional regulations may still apply, data de-identification procedures involve removing, obscuring, or generalizing specific components of a data set so that individuals cannot be re-identified. Proper de-identification must prioritize patient privacy (i.e., the risk of re-identification) while also maximizing the relevancy and quality of the data. Under US authorities, entities can use Safe Harbor, limited dataset, and statistical approaches to de-identifying data. Each of these regimes detail how target data fields are removed or obscured in the de-identification process. However, because of the stringent requirements for data field removal that Safe Harbor requires and the need for signed data use agreements for limited datasets, the statistical method is the most likely method to be considered for real-world data analysis.[56] Because data linkage can increase the risk of re-identification, it is essential that de-identification be performed on the final linked dataset.

Data owners will often mask identifying data by directly removing or anonymizing certain data fields (e.g. date of birth). Demographic and geographic data fields are frequently removed as clinical granularity is added, making characterization of the representativeness of the study population more difficult. The de-identification process also affects traceability and auditability. For some safety purposes, regulators have statutory authority through their public health authority for re-identification and access to PHI.[57] However, that authority does not extend to data regarding other types of regulatory decision-making focused on efficacy. In addition, some datasets available to sponsors and researchers may not have the information necessary to trace back the information to the originating source.

Future sources of RWD will require robust de-identification methods. Free text, such as clinician notes, can be scrubbed of personally identifiable or sensitive data using automated pattern matching and machine learning algorithms, but false negative and false positive rates must be accounted for as they affect the adequacy of de-identification and subsequent extraction algorithms.[58] More commonly, structured information is extracted from free text where false positives may inappropriately include PHI that needs to be removed. Time value data, such as length of stay, time since last visit, and device time stamps, may also need to be masked if it inappropriately increases the probability of re-identification through unique longitudinal signatures.[59] In genetics, research has shown that even a sequence of no more than 30 to 80 independent single nucleotide polymorphisms (SNPs) can uniquely identify a single person.[60]

Thorough documentation on the type of de-identification (deterministic or probabilistic) and re-identification risk thresholds used is critical. Sponsors and researchers need to be confident that de-identification was properly performed and have the information required to clearly assess any impact of de-identification on the data's fitness-for-purpose. Investigators should report which variables were selected for de-identification and what de-identification algorithms were used. Investigators should be clear about known and expected data relevancy tradeoffs due to de-identification transformations.

APPENDIX D: GLOSSARY

Accuracy — Assessment of the validity, reliability, and robustness of a data field.

Common data model — Comprehensive framework that includes definitions, specifications, and operational rules for data to be presented and used in a common manner.

Completeness — Measure of recorded data present within a defined data field and/or data set.

Conformance — Data congruence with standardized types, sizes, and formats.

Consistency — Stability of a data value within a dataset or across linked datasets.

Covariate — Data used to characterize patient populations, balance groups, and/or control for confounding.

Data element/value — A piece of data corresponding to one patient within a data field.

Data field — A technically specified column for data elements.

Data quality — An assessment of the attributes of data needed to answer the question of interest accurately, reliably, and repeatedly.

Data relevancy — Data that are representative of the population of interest and specific to a given clinical and regulatory context.

Data source — A collection of singular or mixed data types whose origin and method of collection are similar (e.g., EHR, claims, PDHA, registry, etc.).

Data type — Data that share a common format or standard data fields (e.g., diagnoses, symptoms, procedure codes, lab results, PROs, home monitor data, etc.).

Exposure — Therapeutic intervention or event under study.

Fit-for-purpose data — An assessment of whether a meaningful, valid, and transparent data set can answer the question of interest given data quality, data relevancy, and the current body of evidence.

Generalizability — The ability of study findings to be externally valid to populations outside of the study.

Historical experience — Data that contains a readily available record of use.

Imputation — Rigorous process for substituting missing values.

Information bias — Systematic distortions in the data under study arising from measurement error.

Longitudinality — Condition of data indexed by time/interval of exposure and outcome time.

Normalization — Adjustments to recorded data for scale or distributional alignment.

Outcome — Key endpoint, diagnostic, test, or result.

Personal digital health applications — Devices and software that capture patient data not well-suited to direct clinical measurement or not captured in natural clinical health care settings.

Personal health information — Identifying and clinical health data protected by government statues, principally the Health Insurance Portability and Accountability Act (HIPAA).

Patient matching — Probabilistic or deterministic algorithms for matching individual patients across data sets.

Plausibility — Recorded values are logically believable given data source and expert opinion.

Provenance — Origin of the data, sometimes including a chronological record of data custodians and transformations.

Randomized controlled trials — Experiment where subjects are randomly selected for treatment to balance observed and unobserved variables.

Real-world data — Data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources.

Real-world evidence — Evidence derived from RWD through the application of research methods. For regulatory applications, RWE can further be defined as clinical evidence regarding the use and potential benefits or risks of a medical product derived from analysis of RWD.

Representativeness — The ability to accurately reflect the characterized population(s) of interest.

Selection bias — Systematic distortions in the representativeness of a sample population under study.

Sufficiency — Statistical and compositional threshold of sample size and data richness needed for accurate analysis.

Systematic bias — Distortions in the data under study due to persistent and non-random causes.

Traceability — Ability to record changes to location, ownership, and values.

Validity — Measure of concordance between a data field and a definitive measure.

REFERENCES

[1] Energy and Commerce Committee. 21st Century Cures Act. Retrieved September 27, 2018, from: https://energycommerce.house.gov/cures

[2] Center for Drug Evaluation and Research. Prescription Drug User Fee Act (PDUFA) - PDUFA VI: Fiscal Years 2018 – 2022. US Food and Drug Administration Home Page. Retrieved September 27, 2018, from: http://www.fda.gov/ForIndustry/UserFees/PrescriptionDrugUserFee/ucm446608.htm

[3] Berger, M. et. al (2017). A Framework for Regulatory Use of Real-World Evidence. Retrieved September 27, 2018, from: https://healthpolicy.duke.edu/sites/default/files/atoms/files/rwe_white_paper_2017.09.06.pdf

[4] Dreyer, N. A. (2018). Advancing a Framework for Regulatory Use of Real-World Evidence: When Real Is Reliable. *Therapeutic innovation & regulatory science*, *52*(3), 362-368.

[5] US Food and Drug Administration. (2013). Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data. Guidance for industry and Food and Drug Administration staff Center for Biologics Evaluation and Research (CBER).

[6] US Food and Drug Administration. (2017). Use of real-world evidence to support regulatory decision-making for medical devices. Guidance for industry and Food and Drug Administration staff. Center for Devices and Radiological Health (CDRH).

[7] Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A. E., Brown, J., Davidson, B. N., ... & Liaw, S. T. (2016). A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. Egems, 4(1).

[8] Weiskopf, N. G., Bakken, S., Hripcsak, G., & Weng, C. (2017). A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. eGEMs (Generating Evidence & Methods to improve patient outcomes), 5(1).

[9] Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A. E., Brown, J., Davidson, B. N., ... & Liaw, S. T. (2016). A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. Egems, 4(1).

[10] Weiskopf, N. G., Bakken, S., Hripcsak, G., & Weng, C. (2017). A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. eGEMs (Generating Evidence & Methods to improve patient outcomes), 5(1).

[11] Thomas, L., & Peterson, E. D. (2012). The value of statistical analysis plans in observational research: defining high-quality research from the start. Jama, 308(8), 773-774.

[12] West, P., Van Kleek, M., Giordano, R., Weal, M., & Shadbolt, N. (2017). Information Quality Challenges of Patient-generated Data in Clinical Practice. Frontiers in Public Health, 5

[13] Wade, R. A., & Krouse, A. T. ABA Health eSource [Internet]. Chicago: American Bar Association; EHRs, upcoding, overpayments, and the False Claims Act–understanding the risks; 2013–2014.

[14] Funk, M. J., & Landi, S. N. (2014). Misclassification in administrative claims data: quantifying the impact on treatment effect estimates. Current epidemiology reports, 1(4), 175-185.e

[15] Dreyer, N. A., Schneeweiss, S., McNeil, B. J., Berger, M. L., Walker, A. M., Ollendorf, D. A., & Gliklich, R. E. (2010). GRACE principles: recognizing high-quality observational studies of comparative effectiveness. The American journal of managed care, 16(6), 467-471.

[16] Epstein, M. (2005). Guidelines for good pharmacoepidemiology practices (GPP). Pharmacoepidemiology and drug safety, 14(8), 589-595.

[17] US Food and Drug Administration. (2007). Guidance for industry: computerized systems used in clinical investigations. Rockville, MD: US Department of Health and Human Services.

[18] US Food and Drug Administration. (2017). Use of real-world evidence to support regulatory decision-making for medical devices. Guidance for industry and Food and Drug Administration staff. 2017.

[19] Van den Broeck, J., Cunningham, S. A., Eeckels, R., & Herbst, K. (2005). Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS medicine, 2(10), e267.

[20] Von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., Vandenbroucke, J. P., & Strobe Initiative. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS medicine, 4(10), e296.

[21] U.S. Food and Drug Administration. (2018). Patient-Focused Drug Development: Collecting Comprehensive and Representative Input Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders

[22] Ketcham, Jonathan D., Laurence C. Baker, and Donna MacIsaac. "Physician practice size and variations in treatments and outcomes: evidence from Medicare patients with AMI." Health Affairs26.1 (2007): 195-205.

[23] Little, R. J., & Rubin, D. B. (2014). Statistical analysis with missing data (Vol. 333). John Wiley & Sons.

[24] Robb, M. A., Racoosin, J. A., Sherman, R. E., Gross, T. P., Ball, R., Reichman, M. E., ... & Woodcock, J. (2012). The US Food and Drug Administration's Sentinel Initiative: expanding the horizons of medical product safety. Pharmacoepidemiology and drug safety, 21, 9-11.

[25] US Food and Drug Administration. (1998). Guidance for industry: E9 statistical principles for clinical trials. Rockville, MD: Food and Drug Administration.

[26] US Food and Drug Administration. (1996). Guideline for Industry Structure and Content of Clinical Study Reports. Center for Drug Evaluation and Research ICH E, 3.

[27] IQVIA. Physician Office Usage of Electronic Health Records Software. May 2018. www.iqvia.com/-/media/iqvia/pdfs/us-location-site/commercial-operations/iqvia-ehr-adoption_2018.pdf. Retrieved September 25, 2018

[28] Tabak, Y. P., Sun, X., Derby, K. G., Kurtz, S. G., & Johannes, R. S. (2010). Development and validation of a disease-specific risk adjustment system using automated clinical data. Health services research, 45(6p1), 1815-1835.

[29] Hripcsak, G., & Albers, D. J. (2012). Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association, 20(1), 117-121.

[30] Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association, 20(1), 144-151

[31] Centers for Medicare & Medicaid Services. (2017, September 19). Hospital Inpatient Quality Reporting Program. Retrieved from https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU.html

[32] Hersh, W. R., Weiner, M. G., Embi, P. J., Logan, J. R., Payne, P. R., Bernstam, E. V., ... & Saltz, J. H. (2013). Caveats for the use of operational electronic health record data in comparative effectiveness research. Medical care, 51(8 0 3), S30.

[33] Kari, F., Bryan, B., & Paul, J. (2009). The use of claims data in healthcare research. The Open Public Health Journal, 2(1).

[34] Safran, C., Bloomrosen, M., Hammond, W. E., Labkoff, S., Markel-Fox, S., Tang, P. C., & Detmer, D. E. (2007). Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. Journal of the American Medical Informatics Association, 14(1), 1-9.

[35] Jollis, James G., et al. "Discordance of databases designed for claims payment versus clinical information systems: implications for outcomes research." Annals of internal medicine 119.8 (1993): 844-850.

[36] Chawla, N., Yabroff, K. R., Mariotto, A., McNeel, T. S., Schrag, D., & Warren, J. L. (2014). Limited validity of diagnosis codes in Medicare claims for identifying cancer metastases and inferring stage. Annals of Epidemiology, 24(9), 666-672.

[37] Sommers, B. D., & Rosenbaum, S. (2011). Issues in health reform: how changes in eligibility may move millions back and forth between Medicaid and insurance exchanges. Health Affairs, 30(2), 228-236.

[38] Sentinel Initiative. (2018, April 18). Health Outcome of Interest Validations and Literature Reviews. Retrieved from https://www.sentinelinitiative.org/sentinel/surveillance-tools/validations-lit-review

[39] US Food and Drug Administration. Clinical outcome assessment qualification program. Retrieved from: http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm284077.htm

[40] Consumer Technology Association. Physical activity monitoring for fitness wearables: Step counting. ANSI/CTA Standard. 2016

[41] Coons, S. J., Kothari, S., Monz, B. U., & Burke, L. B. (2011). The patient-reported outcome (PRO) consortium: filling measurement gaps for PRO end points to support labeling claims. Clinical Pharmacology & Therapeutics, 90(5), 743-748.

[42] Kluetz, P. G., O'Connor, D. J., & Soltys, K. (2018). Incorporating the patient experience into regulatory decision making in the USA, Europe, and Canada. The Lancet Oncology, 19(5), e267-e274.

[43] Lai, A. M., Hsueh, P. Y., Choi, Y. K., & Austin, R. R. (2017). Present and future trends in consumer health informatics and patient-generated health data. Yearbook of medical informatics, 26(01), 152-159.

[44] Petersen, C., & DeMuro, P. (2015). Legal and regulatory considerations associated with use of patient-generated health data from social media and mobile health (mHealth) devices. Applied clinical informatics, 6(1), 16.

[45] U. S. Food and Drug Administration. (2018). Patient-Focused Drug Development: Collecting Comprehensive and Representative Input. Center for Drug Evaluation and Research. Center for Biologics Evaluation and Research.

[46] U. S. Food and Drug Administration. (n.d.). Development Resources - Clinical Outcome Assessment Compendium. Center for Drug Evaluation and Research. Retrieved September 14, 2018, from https://www.fda.gov/Drugs/DevelopmentApprovalProcess/DevelopmentResources/ucm459231.htm

[47] Brajovic S, Blaser DA, Zisk M, Caligtan C, Okun S, Hall M, Pamer CA 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 DOI: 10.2196/medinform.9878

[48] Case, M. A., Burwick, H. A., Volpp, K. G., & Patel, M. S. (2015). Accuracy of smartphone applications and wearable devices for tracking physical activity data. Jama, 313(6), 625-626.

[49] Hsueh, P. Y., Cheung, Y. K., Dey, S., Kim, K. K., Martin-Sanchez, F. J., Petersen, S. K., & Wetter, T. (2017). Added Value from Secondary Use of Person Generated Health Data in Consumer Health Informatics. Yearbook of medical informatics, 26(01), 160-171.

[50] Text mining at AstraZeneca: Comparing real-world evidence to clinical trial events (Tech.). (2018). Retrieved September 27, 2018, from Linguamatics.

[51] Blaser DA, Eaneff S, Loudon-Griffiths J, Roberts S, Phan P, Wicks P, Weatherall J. Comparison of rates of nausea side effects for prescription medications from an online patient community versus medication labels: an exploratory analysis. AAPS Open. 2017. 3:10.

[52] National Center for Health Statistics. (n.d.). National Death Index. Retrieved September 18, 2018, from https://healthdata.gov/dataset/national-death-index

[53] Agarwala, V., Khozin, S., Singal, G., O'Connell, C., Kuk, D., Li, G., ... & Abernethy, A. P. (2018). Real-world evidence in support of precision medicine: clinico-genomic cancer data as a case study. Health Affairs, 37(5), 765-772.

[54] Kheirbek, I., Wheeler, K., Walters, S., Kass, D., & Matte, T. (2013). PM2. 5 and ozone health impacts and disparities in New York City: sensitivity to spatial and temporal resolution. Air Quality, Atmosphere & Health, 6(2), 473-486.

[55] Barash, C. I., Elliston, K. O., Faucett, W. A., Hirsch, J., Naik, G., Rathjen, A., & Wood, G. (2015). Harnessing big data for precision medicine: A panel of experts elucidates the data challenges and proposes key strategic decisions points.

[56] El Emam, K. (2011). Methods for the de-identification of electronic health records for genomic research. Genome medicine, 3(4), 25.

[57] Rosati, K., Evans, B. J., Jorgensen, N., & Soliz, M. (2018, February 1). Sentinel Initiative Principles and Policies, HIPAA and Common Rule Compliance in the Sentinel Initiative (Rep.). Retrieved September 18, 2018, from Sentinel Initative website: https://www.sentinelinitiative.org/sites/default/files/communications/publications-presentations/HIPPA-Common-Rule-Compliance-in-Sentinel-Initiative.pdf

[58] Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S., & Samore, M. H. (2010). Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC medical research methodology, 10(1), 70.

[59] Hripcsak, G., Mirhaji, P., Low, A. F., & Malin, B. A. (2016). Preserving temporal relations in clinical data while maintaining privacy. Journal of the American Medical Informatics Association, 23(6), 1040-1045.

[60] Lin, Z., Owen, A. B., & Altman, R. B. (2004). Genomic research and human subject privacy.