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

Unpacking Real-World Data Curation: Principles and Best Practices to Support Transparency and Quality

Duke-Robert J. Margolis, MD, Center for Health Policy
1201 Pennsylvania Ave, NW Suite 500 • Washington, DC 20004
January 22, 2019

Discussion Guide

Background

In the coming years, there will be emerging opportunities for making better use of real-world data (RWD) within the U.S. Food and Drug Administration's (FDA's) regulatory framework. More consistent use of RWD sources can improve the efficiency of traditional randomized controlled trials, and there may also be opportunities to leverage RWD and resultant real-world evidence (RWE) in support of supplemental approval or labeling actions based on substantial evidence of effectiveness as envisioned in 21st Century Cures and PDUFA VI.

As part of implementation efforts for this legislation, the FDA published a strategic framework to guide the development of a new program for regulatory uses of RWD and RWE.[1] This framework will help FDA and stakeholders explore how RWD/RWE could support pre-market regulatory decisions on medical product effectiveness, specifically evaluating whether RWE could support changes to labeling such as adding or modifying an indication. To guide such decisions FDA will use the following key considerations:

- Is the RWD fit for purpose?
- Does the trial or study design used to generate RWE provide adequate scientific evidence to answer or help answer the regulatory question of interest?
- Did the conduct of the study meet FDA regulatory requirements?

As established in FDA's framework, a fundamental component of developing RWE that can support regulatory decisions will be an appropriate, or "fit-for-purpose," dataset. Ensuring fitness-for-purpose will require addressing important questions related to the underlying RWD. Regulators will need to understand and assess the provenance of RWD that is submitted as part of an evidence package, as well as the curation and transformation steps that were applied as data moved from inputs provided by a source, to raw data output, to development of the analytic file and dataset used for analysis.

There are various ways to describe and track the process of moving data from inputs to a fit-for-purpose dataset. Although a widely accepted framework does not yet exist, this workshop will discuss data curation activities as occurring in a two-stage process (Fig 1).[2] The first stage extracts raw data from different sources and normalizes these data into databases that could be used for general research purposes, otherwise referred to as research-ready data.

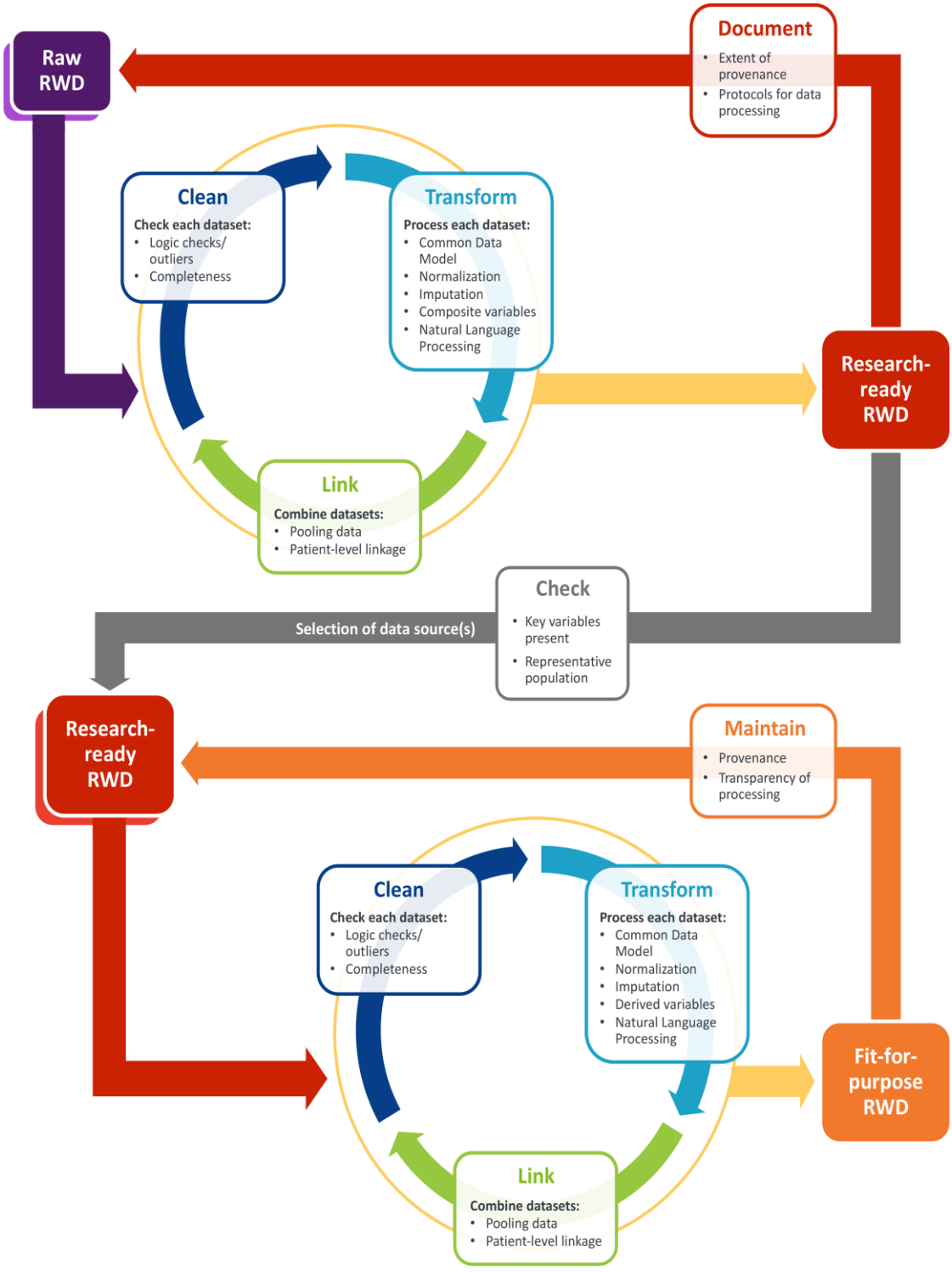

Figure 1. The process of making a fit-for-purpose dataset.

With additional curation steps, selected data from these databases can be further developed into a fit-for-purpose dataset to answer specific study questions, which represent the second stage of data curation.

Across both stages of curation, there are 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). Given the incremental changes made to the data, proper documentation of these steps will help stakeholders assess the accuracy and completeness of the fit-for-purpose dataset.

Due to the growing diversity of source data being collected by different systems and with a variety of data curation practices being implemented, questions remain about how the data curation steps described above might impact the ultimate quality and validity of the RWD. This workshop, convened by the Robert J. Margolis, MD, Center for Health Policy at Duke University under a cooperative agreement with FDA, will gather organizational perspectives and explore potential best practices, where possible, of specific curation steps or techniques. The workshop will provide an opportunity for data curators to come together pre-competitively to 1) aid FDA in better understanding the types of curation characteristics that may need to be documented to support evaluation of the data during regulatory review, and 2) further strengthen the field of large-scale RWD development and curation.

Session I: Transforming Raw Data into Research Ready-Data

This session will cover a range of questions related to the first stage of data curation, which involves cleaning, transforming, and linking raw data from its source into research-ready data (note that considerations of data linkages will be addressed in Session III). Participants will share and consider processes and approaches used to extract, normalize, and map raw data into standardized fields. Discussion will consider the range of basic steps and transformations that stakeholders conduct and how these might apply to different data sources (e.g., lab data, unstructured and structured clinical data, and claims data). As part of the discussion, participants will also consider what validation checks are typically used to evaluate data completeness.

Key Discussion Questions:

- What steps, processes, and technologies are being used for data extraction?
  - What is the experience with automating (e.g., optical character recognition or natural language processing) data extraction processes versus manual practices?
  - How do organizations ensure they are extracting the most recent data from various data sources and how might data access policies play a role in this?
- What are the practices for normalizing raw data into standardized fields using a common data model?
  - What approaches are currently being used for data model mapping to curate research ready data?
  - What are the advantages/disadvantages to mapping raw data into more meaningful clinical concepts at this stage?
  - How might the stage at which data are mapped to clinical concepts within a common data model impact the use of these data for regulatory decision making?
- What are the key considerations for data quality in first stage curation?
  - What biases might be introduced as part of the data extraction and normalization process, and what steps can be taken to mitigate these biases?
  - What are the key considerations to assess data completeness, and validating whether

requested data values were extracted and mapped to the correct fields?
- How are organizations determining when data are missing? What are the underlying reasons for data missingness?

Session II: Study Specific Data Curation to Establish a Fit-for-Purpose Dataset

While research-ready data can support some analyses, with additional curation steps these data can be extracted and transformed into a study-specific dataset that contains the cohort(s) of interest. This second stage of data curation may include additional validation checks beyond data completeness including reconciling conflicting data values (e.g. different values for the same patient recorded on the same day), ensuring values fit within prespecified minimums and maximums, and implementing strategies to address data missingness in order to create a fit-for-purpose dataset. Discussion in this session will consider approaches used to evaluate whether the data are fit-for-purpose, in combination with appropriate study designs and methods, to generate RWE that provides adequate scientific evidence to answer or help answer the regulatory question of interest.

Key Discussion Questions:

- What additional, study specific data curation steps are being performed to develop a fit-for-purpose dataset from research ready data?
  - How are these steps and transformations different from stage 1 curation processes?
  - What techniques are organizations using to clean and reconcile inaccurate or implausible data values to develop the fit-for-purpose dataset?
  - Can proxy variables or data coding techniques be used to fill potential gaps in data completeness to establish the fit-for-purpose dataset?
  - Are there any parameters or protocols that need to be prespecified to prevent biases from being introduced into the fit-for-purpose dataset (e.g. potential challenges with data variance)?
  - How are organizations validating that appropriate data points are being captured to answer the study specific question of interest?

Session III: Linking Multiple Data Sources

This session will consider key linkage considerations for curating data. Specifically, this session will consider the techniques and criteria used to link structured data from claims, EHRs, and non-structured data. This could include key practices and performance measures developed for linking data sources and validating the linking process, as well as strategies aimed at improving common understanding of data terms, concepts, and vocabularies across data sources.

Key Discussion Questions:

- What are the opportunities and challenges of different data linkage methods (pooling versus patient-level linkages)?
  - Are there potential best practices for combining structured data with unstructured data to create a fit-for-purpose dataset? What are the opportunities and challenges with using NLP and machine learning algorithms?

- What are the best practices for patient matching during the linkage process?
  - How can hashing techniques be used to improve the efficiency of patient matching?
  - What levels of confidence are needed for patient matching in research-ready data versus a fit-for-purpose dataset?
  - How could patient level linkages be validated to ensure correctness (e.g., is the right information linked to the right person across the data sources used)?
  - How can protected health information be maintained?
- How does the process of data linkage affect data quality? Are there any dimensions of quality (e.g., characteristics of represented populations) that should be reevaluated in the linked data?

Session IV: Submitting Data Documentation for Traceability and Auditing

As part of regulatory review, it will be important for regulators to understand the quality and characteristics of the data being analyzed and potentially submitted as part of an evidence package. Understanding the provenance of data elements is one key aspect of assessing data quality, and this session will explore how organizations are tracking and documenting provenance as well as other key data quality considerations.

Key Discussion Questions:

- What needs to be documented in Stage 1 compared to Stage 2 data curation?
  - Are there documentation needs specific to each stage of data curation?
  - Do standard documentation, auditing processes, and validation requirements exist for tracing data provenance? Where are the gaps?
  - Are there any privacy laws and regulations that might impact assessment of provenance?
  - What are the most effective technical practices for preserving original values (and medical note changes) for the purpose of provenance?
- What are the potential considerations for assessing data documentation?
  - How could documentation for data curation build on key priorities identified in FDA's RWE Program to assess data standards and implementation strategies for using RWD and RWE?
  - What would a standard format look like for a fit-for-purpose dataset submitted as part of an evidence package to FDA? Could submission requirements include HL7's FHIR standard, or standards developed under the Clinical Data Interchange Standards Consortium?
  - What are common features of highly curated research-ready data that could support more efficient adjudication of a fit-for-purpose dataset?

Funding for this conference was made possible in part by a cooperative agreement from the U.S. Food and Drug Administration Center for Drug Evaluation and Research. The views expressed in written conference materials or publications and by speakers and moderators do not necessarily reflect the official policies of the Department of Health and Human Services nor does mention of trade names, commercial practices, or organizations imply endorsements by the U.S. Government.

Appendix A: Suggested Background Reading

Curtis, L.H., Weiner, M.G., Boudreau, D.M., Cooper, W.O., Daniel, G.W., Nair, V.P., Raebel, M.A., Beaulieu, N.U., Rosofsky, R., Woodworth, T.S., & Brown, J.S. (2012). Design considerations, architecture, and use of the Mini-Sentinel distributed data system. *Pharmacoepidemiology and Drug Safety, 21*(1), 23-31. https://doi.org/10.1002/pds.2336.

Curtis, M.D., Griffith, S.D., Tucker, M., Taylor, M.D., Capra, W.B., Carrigan, G., Holzman, B., Torres, A.Z., You, P., Arnieri, B., & Abernethy, A.P. (2018). Development and Validation of a High-Quality Composite Real-World Mortality Endpoint. *Health Services Research, 53*:6, 4460-4476. https://www.ncbi.nlm.nih.gov/pubmed/29756355

Daniel, G., Silcox, C., Bryan, J., McClellan, M., Romine, M., Frank, K. (2018). Characterizing RWD Quality and Relevancy for Regulatory Purposes. Retrieved from: https://healthpolicy.duke.edu/sites/default/files/atoms/files/characterizing_rwd.pdf

Dusetzina SB, Tyree S, Meyer AM, et al. Chapter 4: An Overview of Record Linkage Methods. Linking Data for Health Services Research: A Framework and Instructional Guide. Rockville (MD): Agency for Healthcare Research and Quality (US); 2014 Sep. https://www.ncbi.nlm.nih.gov/books/NBK253312/.

Kahn, M.G., Brown, J.S., Chun, A.T., Davidson, B.N., Meeker, D., Ryan, P.B., Schilling, L.M., Weiskopf, N.G., Williams, A.E., & Zozus, M.N. (2015). Transparent Reporting of Data Quality in Distributed Data Networks. *eGEMS (Generating Evidence & Methods to Improve Patient Outcomes*,3(1):1052. doi: 10.13063/2327-9214.1052

Raebel, M.A., Haynes, K., Woodworth, T.S., Saylor, G., Cavagnaro, E., Coughlin, K.O., Curtis, L.H., Weiner, M.G., Archdeacon, P., & Brown, J.S. (2014). Electronic clinical laboratory test results data tables: lessons from Mini-Sentinel. *Pharmacoepidemiology and Drug Safety, 23*(6), 609-619. https://onlinelibrary.wiley.com/doi/full/10.1002/pds.3580.

Kalilani, L., Halpern, R., Seare, J., & Dedeken, P. (2018). The challenges of assessing effectiveness of lacosamide using electronic medical record databases. *Epilepsy & Behavior,85*, 195-199. https://www.sciencedirect.com/science/article/pii/S1525505018300726.

Nunes, A.P., Yang, J., Radican, L., Engel, S.S., Kurtyka, K., Tunceli, K., Yu, S., Iglay, K., Doherty, M.C., & Dore, D.D. (2016). Assessing occurrence of hypoglycemia and its severity from electronic health records of patients with type 2 diabetes mellitus. *Diabetes Research and Clinical Practice, 121*, 192-203. https://www.ncbi.nlm.nih.gov/pubmed/27744128

Qualls, L.G., Phillips, T.A., Hammil, B.G., Topping, J., Louzao, D.M., Brown, J.S., Curtis, L.H., & Marsolo, K. (2018). Evaluating Foundational Data Quality in the National Patient-Centered Clinical

Research Network (PCORnet®). *eGEMS (Generating Evidence & Methods to Improve Patient Outcomes*, *6*(1):3, 1-9. http://doi.org/10.5334/egems.199

Reich, C., Ryan, P.B., Stang, P.E., & Rocca, M. (2012). Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. *Journal of Biomedical Informatics, 45*(4), 689-696. https://doi.org/10.1016/j.jbi.2012.05.002

U.S. Food and Drug Administration. (2018). Framework for FDA's Real-World Evidence Program. Retrieved from: https://www.fda.gov/downloads/ScienceResearch/SpecialTopics/RealWorldEvidence/UCM627769.pdf

Whitman, E.D., Liu, F.X., Cao, X., Diede, S.J., Haiderali, A., & Abernethy, A.P. (2018). Treatment patterns and outcomes for patients with advanced melanoma in US oncology clinical practices. *Future Oncology*. https://doi.org/10.2217/fon-2018-0620