PHUSE EU Connect 2023

Paper RW03

Submission Standards for Real World Data: Gaps, Limitations and Recommendations

Jeff Abolafia, P21, Blue Bell, PA, USA
Sarah Ferko, IBM Consulting, USA
Ingeborg Holt, IBM Consulting, USA

ABSTRACT

There is a rapid increase in the use of RWD to support marketing applications. In 2020 over 75% of NDAs and BLAs submitted to FDA included a RWD study to support safety and/or efficacy claims. However, submission standards and related documentation required for RWD are still in development and not well understood. This can make both regulatory submission and review more difficult. As a result, sponsors must develop an appropriate strategy for submitting RWD.

This presentation follows up on 2 previous RWD papers presented at recent PhUSE conferences. Specifically, this presentation focuses on traceability and documentation for marketing applications using RWD. We examine the challenges and limitations of using current standards. Next, we present recommendations and solutions for submitting studies containing RWD, based on FDA guidance and recent marketing applications containing RWD. These recommendations are intended to help sponsors submitting RWD enhance the traceability and documentation required for regulatory review.

INTRODUCTION

In the past several years there has been a sharp increase in the number of NDAs/BLAs that include real world data (RWD). While RWD can be used to provide regulatory authorities with additional evidence of efficacy or safety, it currently presents numerous challenges to regulatory reviewers (and sponsors). In our previous two papers (1,2), we examined the current regulatory landscape for submitting RWD, the gaps and limitations for submitting RWD using the current data standards and provide recommendations for submitting RWD using the current standards. In this paper, we again look at the challenges submitting RWD. However, this paper focuses on the traceability and documentation necessary for regulatory review. We examine the challenges for providing traceability when submitting RWD to support a marketing application and then present several recommendations to facilitate regulatory review. Our recommendations are based on FDA guidance and recent marketing applications containing RWD. These recommendations are intended to help sponsors submit RWD that are more compliant with current submission standards and enhance the traceability and documentation required for regulatory review.

BACKGROUND

REAL WORLD EVIDENCE (RWE)

RWE can come from a variety of sources including EHRs, payer administration claims, digital health technology devices (DHTs), and patient registries. Sponsors are submitting an increasing amount of RWD to support marketing applications and this trend is likely to accelerate. Aetion (3) estimates that approximately 78% of approved NDAs and BLAs in 2020 included an RWD study to provide evidence of safety and/or efficacy compared to 53% in 2019.

Section 505F(b) of the FD&C Act defines RWE as “data regarding the usage, or the potential benefits or risks, of a drug derived from sources other than traditional clinical trials” (21 U.S.C. 355g(b)). In developing its RWE program, FDA believes it is helpful to distinguish between the sources of RWD and the evidence derived from that data. Evaluating RWE in the context of regulatory decision-making depends not only on the evaluation of the methodologies used to generate the evidence but also on the reliability and relevance of the underlying RWD; these constructs may raise different types of considerations. For the purposes of this framework, FDA defines RWD and RWE as follows:

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 the clinical evidence about the usage and potential benefits or risks of a medical product derived from analysis of RWD.

For the purposes of this paper, we will use the working definition of RWD as data relating to patient health status and/or the delivery of health care that is not collected under a protocol.

REGULATORY LANDSCAPE

The 21st Century Cures Act (4) enacted on December 13, 2016, mandated the FDA to create a framework to evaluate RWE to support marketing applications for new drugs or new indications for drugs already approved. Subsequently, in December 2018, the FDA published its framework for RWE: "Framework for FDA'S Real World Evidence Program"(5). The framework focused on three areas: whether RWD are fit for use, whether the study design can provide adequate evidence, and whether the study conduct meets regulatory requirements. The FDA will focus on these three areas when evaluating RWD submitted in support of a marketing application.

In 2021 the FDA followed up, issuing four more guidance that provided more detail as well as implementation strategies for using RWD to support marketing applications for drugs or biological products. The first guidance, "Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision Making for Drug and Biological Products" (6) focused on validation of definitions for study design elements; and data provenance and quality throughout the study lifecycle.

In October 2021 the FDA issued a second guidance: "Data Standards for Drug and Biological Product Submissions Containing Real-World Data Guidance for Industry" (7). This guidance outlined data standards required when submitting RWD in support of a marketing application. The takeaway message is that RWD must be submitted using the standards documented in the FDA Data Standards Catalog (8). For now, that means RWD must be submitted using CDISC standards. The guidance also provides advice on mapping RWD to CDISC format, and the documentation required. With respect to RWD access, the guidance advises that sponsors should confirm that they are able to submit patient-level data for any RWD used in studies in support of a marketing application required under a New Drug Application and Biologics License Application. The guidance also goes on to say that sponsors will need to convert real-world data (RWD) sets into established standards (CDISC) when submitting this information in support of a marketing application. While the agency acknowledges that its current catalog of standards doesn't necessarily reflect data derived from real-world sources, it indicated that it's looking into updates – and recommended mapping, for now.

TRACEABILITY

Traceability can be defined as:

> "Traceability is built by clearly establishing the path between an element and its immediate predecessor. The full path is traced by going from one element to its predecessors, then on to their predecessors, and so on, back to the SDTM datasets, and ultimately to the data collection instrument." (9)

The guidance issued in 2021 emphasize the importance of submitting patient level data, validation of definitions for study design elements, and data provenance throughout the study lifecycle. The "Study Data Technical Conformance Guide" (10) outlines the regulatory need:

> "An important component of a regulatory review is an understanding of the provenance of the data (e.g., traceability of the sponsor's results back to the CRF data). Traceability permits an understanding of the relationships between the analysis results (tables, listings and figures in the study report), analysis datasets, tabulation datasets, and source data."

I addition, FDA guidance "Data Standards for Drug and Biological Product Submissions Containing Real-World Data Guidance for Industry" (7) stresses the regulatory need for traceability:

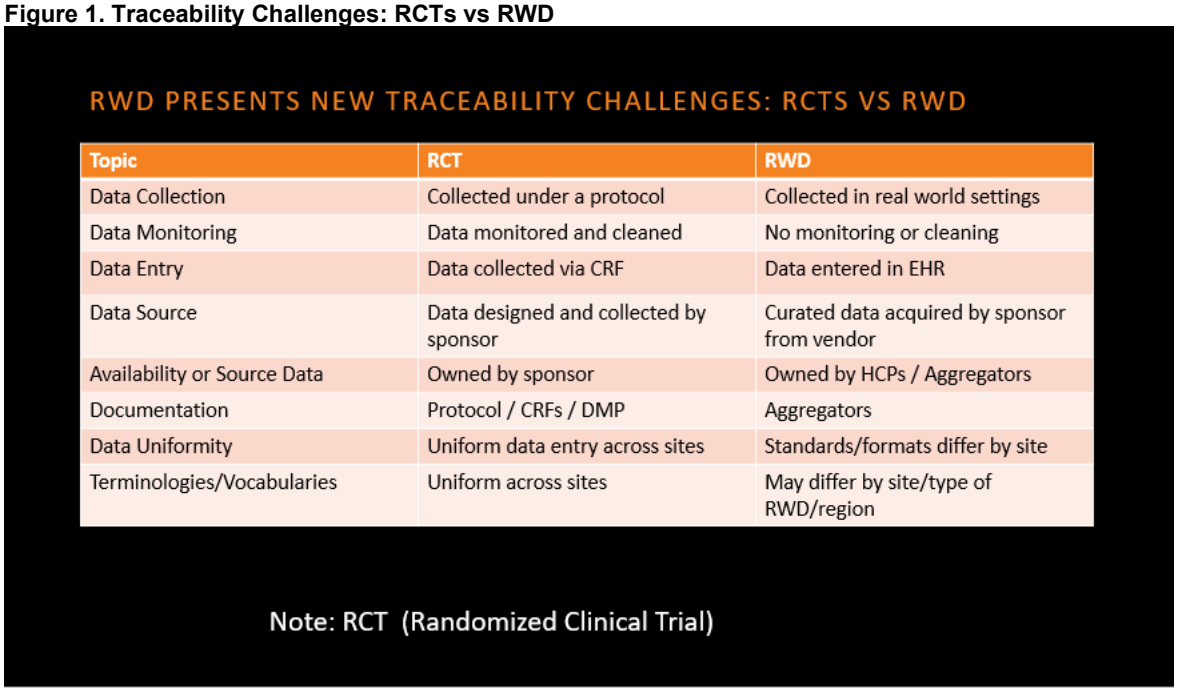
> "Sponsor should include a data dictionary that documents the definition of every data element used and all relevant information about the element, such as its relationships to other data, origin, usage, and format."

In summary FDA guidance dictates that the level of traceability for RWD is equal to that of randomized clinical trials for regulatory review and to support a marketing application.

Traceability Challenges when submitting RWD

The landscape for submitting RWD is still evolving. FDA guidance cited above indicates that the data standards and documentation required for submitting RWD are the same as for submitting RCT data. This presents numerous challenges for sponsors. These challenges are summarized in Figure 1. Most of these challenges stem from the fact that RWD, by definition, is not collected under supervision of a protocol and research study staff.

Figure 1. Traceability Challenges: RCTs vs RWD



Challenge # 1 RWD Not Collected Under a Protocol

While RCT data is collected following a process outlined in a protocol and other related study documents, RWD is collected in real world settings. The protocol and Data management plan (DMP) provide a blueprint for how a research study will be conducted. This includes what assessments will be performed, when they will be performed, and how they will be administered. The DMP presents more granular details. It includes a copy of the annotated CRF, each of the data elements and their attributes, and edit checks for each data element. The DMP also outlines the data monitoring and data cleaning plan. RWD is not collected for a clinical study and therefore is not collected under a protocol or a DMP.

Challenge # 2 RWD Not Monitored or Cleaned

RCT is entered by trained site personnel and closely monitored from the point of data collection through data lock by trained clinical data monitors. Any edits or changes to the source data are thoroughly documented and available for inspection by regulatory authorities. RWD is not monitored or cleaned.

Challenge # 3 Data Source

RCT data comes from systems designed and collected by the sponsor (or agents of the sponsor). As a result, sponsors control the entire process of data collection and transformation from source to submission. On the other hand, most RWD is not directly collected by the sponsor, it is collected in real world settings as patients interact with various health care settings and clinical staff. In many cases RWD is acquired from vendors who have already curated the source data. This presents numerous challenges for documenting how the data submitted was created from the source data.

Challenge # 4 Source Data Availability

The FDA guidance "Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision Making for Drug and Biological Products" (3) states that with respect to RWD access, FDA advises

that sponsors should confirm that they are able to submit patient-level data for any RWD used in studies in support of a marketing application required under a New Drug Application and Biologics License Application. For RCT studies, sponsors have control of the data from source to submission. Submitting patient level data is usually not an issue. For RWD, sponsors rarely control the provenance of data from collection to submission. In most cases RWD is acquired from a RWD vendor and in some cases only aggregated data is available.

Challenge # 5 Data Uniformity

In RCTs data is collected at sites using a standard CRF composed of standard data elements. Site personnel are trained to perform assessments and enter data in a uniform manner. In addition, all sites enter data into a single EDC or eCOA application. For RWD, there is no expectation of uniformity or standardization across health care EHR systems. Even for a given vendor, EHR systems are not implemented in a standard way across health care providers. Underlying databases and the EHR interfaces are customized to meet the needs of the healthcare provider using them. Uniform standards for submission of the data exist (HL7 V2 and moving to HL7 FHIR) but there is no review of how the underlying database is mapped to the exchange information.

Challenge # 6 Terminology

There is a lack of harmonization in terminology between clinical research and healthcare, and coding practices differ both between types of RWD and within types of RWD. Many concepts have different names and/or meanings, and different corresponding controlled terminology. For example, within RWD, SNOMED is used to code problems in an EHR Problem List (as mandated by ONC), but International Classifications of Diseases (ICD) coding is used for diagnoses in EHRs and claims data. Medications may be coded using RxNORM or NDC, and laboratory tests are coded using Logical Observation Identifiers Names and Codes (LOINC). Conversely, RCTs use the same terminology and coding systems across sites and studies. For example, RCTs use MedDRA and WHODrug to code adverse events and medications, respectively. CDISC controlled terminology is also leveraged across other domain areas. Recoding RWD terminology can make source data verification a challenge in many cases since current submission standards (CDISC) do not contain standard data elements to represent both source and submission terminology or coding systems.

Challenge # 7 Traceability when Aggregating Longitudinal DHT Data

Due to advances in technology, data collected from Digital Health Technologies are increasingly being captured to support marketing applications. Digital Health Technologies (DHTs) are systems that use computing platforms, connectivity, software, and/or sensors for health care and related uses (13,14). They include technologies intended for use as a medical product, in a medical product, or as an adjunct to other medical products (devices, drugs, and biologics). DHTs often consist of sensor hardware that allows for continuous or intermittent recording of physiological and/or behavioral data (e.g., blood pressure, physical activity, glucose levels). DHT data includes wearables and other types of data and is one of the many types of RWD according to FDA's definition of the types of data that comprise RWD. Sensor-based DHTs used in remote data acquisition can measure longitudinal data over an extended period (e.g., days, weeks, months). For example, activity trackers or continuous glucose monitors can be used to capture continuous, granular measurements over time. When summarizing these longitudinal measurements, traceability can be a challenge when comprehensive documentation is not provided for each 'level' of summarized data. Further, the data output by the device and what is supplied by the device vendor can vary by devices, and device algorithms may be proprietary.

RECOMMENDATIONS

Communication

Early communication with regulatory authorities and vendors is critical for a successful RWE submission. FDA Guidance "Considerations for the Use of Real-World Data and Real World Evidence to Support Regulatory Decision-Making for Drug and Biological Products Guidance for Industry" (11) issued in August 2023 states:

"Sponsors planning to use a non-interventional study to support a marketing application should engage with FDA early in the drug development process using an appropriate regulatory pathway (e.g., requesting a Type C meeting through an existing IND for the product)."

When requesting this meeting sponsors should provide a draft protocol and statistical analysis plan (SAP). These documents should specify the proposed study design and proposed data sources to answer the research question of interest. This will allow FDA to evaluate proposed data sources to ensure they are fit for purpose and are not biased to favor a certain conclusion. The guidance also advises sponsors to discuss access to the RWD with FDA. Sponsors should ensure they can submit patient level data (or they can arrange for a third party to provide access) that is used for analysis in a marketing application.

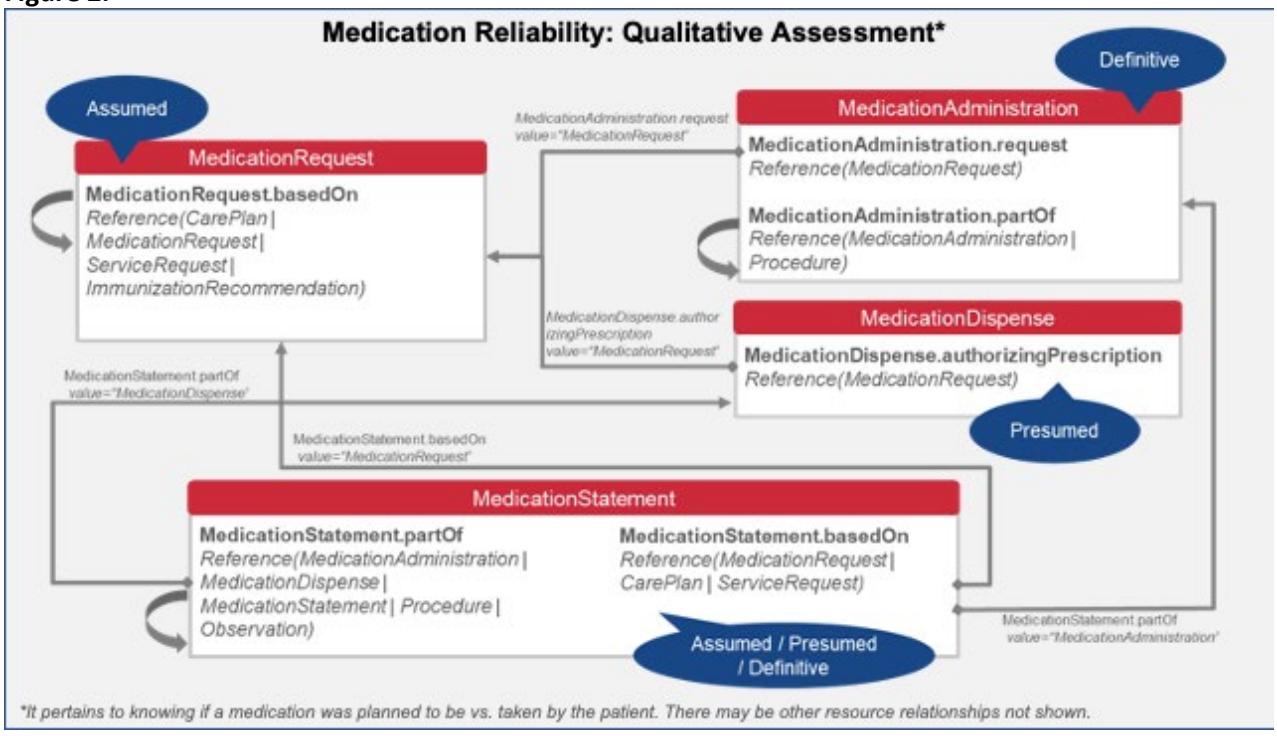
Communication with RWD vendors is also critical when submitting RWD. In short, make sure you know what you are buying. When possible, make sure to vet the data before purchasing it. Some considerations are: 1) will patient level data be available for submission; 2) after exclusions, what will be the sample size; 3) are data for key variables and covariates available; 4) what is the amount of missing data for key variables and covariates; and 5) is the information on data provenance and traceability of data elements to the source available and adequate for a regulatory submission.

Multiple Data Sources for Conformation

In RCTs, the protocol specifies how an assessment is conducted, who performs the assessment, and where the assessment is completed. As a result, this information is usually not captured during data collection, nor has it been required by reviewers. Since RWD data is not collected under a protocol, reviewers would like more information surrounding the context of an assessment to increase confidence and understanding of the data. Where applicable, we recommend adding supplemental "contextual" variables as additional qualifiers to both SDTM and ADaM datasets. These include, who conducted the assessment, the name and location of the site where the assessment was done, and how the assessment was performed. The rationale for providing these contextual variables is that all sites are not equal (large medical center versus corner drug store), clinical staff have different skill sets and specializations, and reviewers are interested in the method(s) used to perform a given assessment.

A second way to increase confidence in RWD is to include multiple data sources for questionable data. Exposure and death data are two use cases where this strategy should be employed. In EHRs, exposure to treatment data typically resides in a variety of places or record types. These include: 1) medication request or medication ordered records; 2) medication dispensed records; 3) medication administered records (medication administered at a clinical site); and medication statement records (medication usage reported by a patient or significant other). EHRs may or may not contain medication data from one or more sources. As a result, it may be difficult to construct exposure from the sources outlined above. Data is usually not available from all relevant record types and different record types may contain conflicting information. Figure 2 from the HL7 implementation guide "Retrieval of Real World Data for Clinical Research" (12) illustrates the various sources of medication information as well as the reliability of each source.

Figure 2.



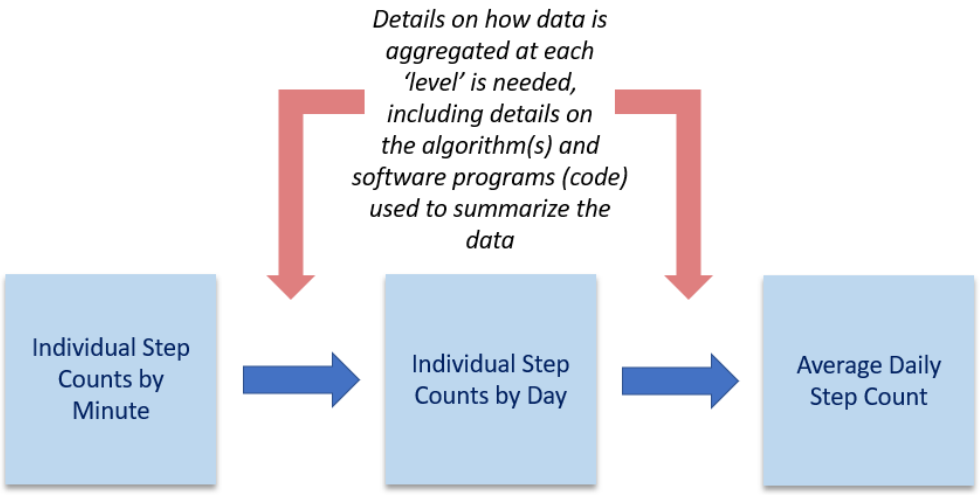
We recommend submitting the collected data from all available sources in a format as "SDTM" as possible. This involves either creating a custom exposure domain or adding as many supplemental variables as need to the SDTM EC and EX domains. A data element should be added which documents the source of each record type. In ADaM exposure to a given treatment can be constructed from all available sources as specified in the statistical analysis plan. Furthermore, a derived "confidence" variable can be useful to report the level of confidence in the accuracy of exposure data.

A second use case is death data. Unless a patient dies in a hospital, death data may not be captured (or captured accurately) in an EHR. Questions may arise around the accuracy of patient death data, including death date and cause of death. This is problematic as valid information about data and cause of death are often critical for the evaluation of efficacy or safety for a drug or biologic. If patient death data is missing or questionable, we recommend aggregating additional death data from multiple sources such as state death records, the National Death Index (NDI), or the Social Security Death Index (SSDI). As with exposure data, variables can be added to report the source of death data and the level of confidence in the accuracy of the death date and cause of death.

Traceability when Aggregating Longitudinal DHT Data

Longitudinal data that is captured and summarized based on the selected DHT-derived endpoint introduces the need for reporting comprehensive traceability for each 'level' of data (i.e., each summarization). The full data flow starts with data output from the device through to the final endpoint data. The example below in Figure 3 shows a step count example and only highlights three 'levels' of data, from minute-level step counts to a summary measure that calculates average daily step count for each subject. When data is aggregated, comprehensive traceability documentation should be provided for each 'level'. It is recommended that a data flow diagram like Figure 3. is included in the cSDRG, as well as a detailed description of the data flow from source to submission.

Figure 3. Traceability Flow Diagram



Long Term Considerations

CDISC standards like SDTM, the Therapeutic Areas Guides (TAUGs), and define.xml are designed for Randomized Clinical Trials (RCTs) or intervention studies which are carried out in a highly controlled environment with a physician and staff trained on the protocol. Data entry is done using an eCRF on which the staff is trained and is scrupulously monitored by automated edit checks and the sponsor. This is in stark contrast to the majority of EHR data, which is collected as part of the patient's care, where the primary goal of documentation is for transactions (e.g., prescription sent to pharmacy) or for billing purposes. Initially it was thought that the sheer volume of RWD would make up for the lack of controlled schedule and monitored data entry, but in the majority of RWD submissions, the number of patients has been fewer than 200 and thus each data point has the potential to affect the statistics. As we move into the next phase of EHRs, from adoption into optimization, hopefully some of these issues will be solved or mitigated.

In addition to the difference in data collection, the variables collected in SDTM are not designed to analyze RWD. Site and investigator information is collected outside of SDTM, as this is not expected to change, and only the site number and investigator id is preserved in SDTM's DM table. If a subject moves and changes sites, this is handled as a major change within a clinical trial. In contrast, in RWD the patient will likely be treated at many different sites of different types (emergency room, specialty doctor, primary care, travel doctor, minute clinic, telehealth) by many different healthcare providers with some of that information making it back to the primary care physician, but likely not all of it. When assessing the validity of a diagnosis or understanding a treatment, this information becomes relevant and is needed by the reviewer to get a complete picture of the patient's journey.

Added to the complication is repackaging RWD into SDTM, although some of the variables are there, they aren't there in the right place, such as the need to identify the treating physician for each patient visit, some are not, such as the

specialty. It may be important to know if a Stevens Johnson Syndrome diagnosis was made by a primary care physician or a dermatologist or emergency room physician. This information can be added into the Supplemental Qualifier domain(s) but there is currently no support for analyzing these variables statistically in most tools designed to work with SDTM. In addition, without controlled vocabulary and variable naming, each analysis will need to be customized adding to the review time and effort. Identification and standardization of these variables, even if only for the SUPPQUAL domain, identifying and standardizing core variables is a useful first step and has precedence in CDISC with the treatment emergent SUPPQUAL variable.

Finding a way to represent data certainty is another important factor in the use of RWD. Data linking solutions are cumbersome and outdated and hampered (rightly or wrongly) by PHI restriction. If the time consuming and expensive work of linking EHR and claims data is done, there is currently no way to represent the added certainty that linking this data can support. In addition, if multiple records in the EHR are used to confirm drug prescription (e.g., medication order and medication dispense records are matched), this should be able to be recorded with more certainty than just a medication order record. The current CDISC SDTM EC and EX domains may be repurposed for this but uniformity about how this is carried out would be useful for the community.

To address the concern of traceability from the source of collection for RWD, one solution is for the FDA to require annotated EHR or EHRs be provided along with the data. This would function in the same manner as an annotated CRF to provide context for how the data was entered into the form and show the reviewer what other choices could have been made. In a case where RWD is collected from EHRs from 2 or more healthcare systems, it would also provide some clarity on how the data might differ between the 2 systems and thus help mitigate bias. For example, an annotated EHR should show the reviewer if a particular data element was collected using controlled terminology or free text. Submitting source documentation like an annotated EHR (or claims form) is currently not required for RWD and could present challenges, particularly if, data is coming from a 3rd party provider and not from the health system directly. However, if phased in over time, 3rd party providers could meet this requirement as they likely do go through a mapping step for each EHR system ingested and should have this information available.

As we move forward with adoption of EHRs and data standards, there is a natural tension between the desire to innovate and the interoperability that standardization helps to promote. Balancing these two concepts will be important as we move forward into a more electronic world and more sources of data and more factors that influence patient health will likely be found. Any data standard chosen should be flexible enough to allow new elements to be incorporated quickly. This was seen in the recent pandemic where it became useful to collect a person's profession to understand their exposure risk. Focusing on the outcome of usefulness to public health should be used as the primary principle guiding the choice, development, and/or enhancement of data standards.

CONCLUSIONS

Real World Data and the evidence that can be derived from it holds the promise of bringing safe, effective drugs to market more quickly. However, given the potential for bias and the variable quality of RWD as it is currently collected in the healthcare system, the FDA must be able to understand the shortcomings of the data to be able to use it meaningfully for decision making. This paper has discussed several ways to mitigate these risks, including identifying core variables for RWD, providing traceability through an EHR, and a mechanism to represent using secondary sources and other information to qualify the validity of the data. As the US moves past its initial stage of adoption of EHRs and other electronic sources of data, it is the hope of the authors of this paper and others, that increased attention (and technology) can be focused on the collection of data during routine healthcare such that this data may become "regulatory grade" data. Currently there are modules being designed in HL7 FHIR to help accommodate clinical trial data, but ideally all patient data will be carefully collected and recorded such that it can be reused to conduct trials not yet designed and far in the future. Improved collection will help both patients and physicians as they seek to optimize patient care and use the data contained in RWD to understand the treatments that produce the best outcomes.

REFERENCES

(1) Abolafia, J, Ferko, S, & Holt, I. (2022). "Submission Standards for RWD: The Good, the Bad and the Ugly". Paper presented at the PHUSE Annual Conference 2022, Belfast, United Kingdom. https://phuse.s3.eu-central-1.amazonaws.com/Archive/2022/Connect/EU/Belfast/PRE_RE09.pdf
(2) Ferko, S., Holt, I., & Abolafia, J., (2023). "Challenges and Considerations for Submitting Real World Data". Paper presented at the PHUSE US Annual Conference 2023, Orlando, FL. https://phuse.s3.eu-central-1.amazonaws.com/Archive/2023/Connect/US/Florida/PRE_RE05.pdf
(3) Aetion. "Aetion_eBook_2021_The role of RWE in FDA Approvals". https://aetion.com/evidence-hub/welcome-to-the-evidence-hub/
(4) 21st Century Cures Act. (2016). https://www.congress.gov/114/plaws/publ255/PLAW-114publ255.pdf
(5) FDA Guidance (2018). "Framework for FDA's Real-World Evidence Program". December 2018.

https://www.fda.gov/media/120060/download
(6) FDA Guidance (2021). "Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision Making for Drug and Biological Products". September 2021. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory
(7) FDA Guidance. (2021). "Data Standards for Drug and Biological Product Submissions Containing Real-World Data Guidance for Industry". October 2021. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/data-standards-drug-and-biological-product-submissions-containing-real-world-data
(8) FDA Guidance. (2023). "Data Standards Catalog". April 2023. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/data-standards-catalog
(9) CDISC. 2022. "ADaM Examples of Traceability v1.0". May 2022. https://www.cdisc.org/standards/foundational/adam/adam-examples-traceability-v1-0
(10) FDA Guidance. (2023). "Study Data Technical Conformance Guide – Technical Specifications Document". June 2023. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/data-standards-catalog
(11) FDA Guidance. (2023). "Considerations for the Use of Real-World Data and Real World Evidence to Support Regulatory Decision-Making for Drug and Biological Products Guidance for Industry". August 2023. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-real-world-data-and-real-world-evidence-support-regulatory-decision-making-drug
(12) HL7 Implementation Guide (2023). "Retrieval of Real World Data for Clinical Research". May 2023. https://build.fhir.org/ig/HL7/vulcan-rwd/medications.html
(13) FDA Guidance. (2023). "Framework for the Use of Digital Health Technologies in Drug and Biological Product Development". March 2023. Framework for the Use of DHTs in Drug and Biological Product Development (fda.gov)
(14) FDA Guidance. (2021). "Digital Health Technologies for Remote Data Acquisition in Clinical Investigations". December 2021. Guidance for Industry (fda.gov)

CONTACT INFORMATION

Jeffrey Abolafia
Certara/Pinnacle 21
mailto:jabolafia@pinnacle21.com