Proposed Contents of a Real-World Data Quality and Relevance Package for Regulatory Submissions

Cynthia J. Girman[a,1], Christina Mack[b], Rachel Leahy[a,c], Dana Teltsch[d], Morgan Marks[e], Grazyna Lieberman[f], Molly L Aldridge[g], Christopher Atzinger[h]

[a] CERobs Consulting, Mills River, NC USA

[b] Real World Solutions, IQVIA, Durham, NC USA

[c] Paradigm Bioscience, LLC, Annapolis, MD USA

[d] Takeda, Cambridge, MA USA

[e] Moderna, Cambridge, MA USA

[f] Retired

[g] CERobs Consulting, Mills River, NC USA

[h] Astellas, Northbrook, IL USA

ABSTRACT

Real world data is increasingly used for regulatory decisions. Based on regulatory guidance documents on data quality and relevancy for real world data, TransCelerate released considerations for audit readiness in 2023. However, the content and level of detail for the actual documentation needed for real world evidence submissions to the Food & Drug Administration are unclear. A group of leaders from the International Society for Pharmacoepidemiology, the Professional Society for Health Economics and Outcomes and Research, and TransCelerate leaders convened to develop a tangible view of a data package justifying real world data relevance and data quality for implementing a real world evidence protocol. Depending on specific data needed for the research question and regulatory decision, each real world data source - evaluated for feasibility - must be described, including the type of data, its accuracy and completeness, adequacy in capturing study variables, details of the data processing and extraction, transformation and loading documentation, as well as audit trails and ideally, traceability to the source. A dynamic framework was developed to highlight pertinent details needed for regulators to fully review a real world evidence submission from a data quality and relevancy viewpoint. These details should help researchers understand the documentation needed to justify the evaluation and selection of real world data sources to address a specific research question and regulatory decision. In addition, data providers can use it to clarify what their customers need for regulatory submissions including real world data, and to remind regulators what documentation they may wish to request.

[1] Corresponding Author: Cynthia J Girman Email: : Cindy.girman@cerobs.com

Keywords: Real world data, data quality, regulatory submissions, data relevancy, health technology assessment

1. Introduction

Since early 20th century, the gold standard for evidence on medical interventions are randomized controlled clinical trials (RCTs) with sufficient sample size balance and intervention groups on both measured and unmeasured factors. A standardized RCT protocol carefully defines the study population and controls how and when measurements are taken and allowable concomitant therapies. This reduces variability so that treatment effects can be detected, if they exist. However, the degree of control and the highly selected participants typical of RCTs can yield results that may not be translatable to all patients who use a medicinal product (Kennedy-Martin et al., 2015). In addition, typical RCTs are inefficient, take inordinate time to conduct and are prohibitively costly (Rodriguez et al., 2019).

Electronic health records (EHR) are becoming nearly universal among practicing physicians. Real world data (RWD) from insurance claims and EHR reflect routine clinical practice and have been used for decades to further study post-marketing product safety. Recently, regulators have issued guidance documents on the use of real world evidence (RWE) for regulatory decisions, that focus on whether RWD holds relevance for the research question and are of acceptable quality (Berger et al., 2017; United States Food & Drug Administration [FDA], 2018; FDA, 2021; FDA 2023; European Medicines Agency [EMA], 2023a). European Medicines Agency (EMA) issued a data quality framework for medicines regulation in 2023, focusing on completeness and coverage, coherence, plausibility, traceability and timeliness, in addition to relevance to the research question (EMA, 2023b). Hence, relevance and data quality of RWD sources are key concerns for the use of RWD for US and EU regulatory decisions.

TransCelerate, a consortium of industry and data provider partners, focused on audit readiness to help sponsors and data providers understand regulator expectations and needed documentation for data quality (TransCelerate Biopharma, 2023). Their considerations outlined a practical, relevant roadmap for developing documentation on data quality and relevancy intended for regulators, but did not give a full picture of what such a data package might look like. We created a dynamic framework for a data package to help delineate contents and level of detail expected, which we believe will be useful for preparing submission materials and to guide early discussions with data providers and regulators.
Unlike a traditional research paper, this paper introduces a proposed tool that could be useful for RWE submissions to help ensure that appropriate documentation is provided.

2. Approach

The TransCelerate RWD Audit Readiness initiative was formed in 2020 and used papers and guidance documents from Duke-Margolis and the U.S. Food & Drug Administration (FDA) on the use of RWE in regulatory decisions about products (TransCelerate Biopharma, 2023). The objective was to generate a list of considerations related to RWD quality to facilitate data source selection and support

future regulatory audits (Figure 1) (TransCelerate Biopharma, 2023). Their approach included a comprehensive literature review, identification of conceptual pillars, stakeholder feedback and a survey of data providers, allowing drafting of considerations for each conceptual pillar, followed by internal review and revising according to public comments (TransCelerate Biopharma, 2023).

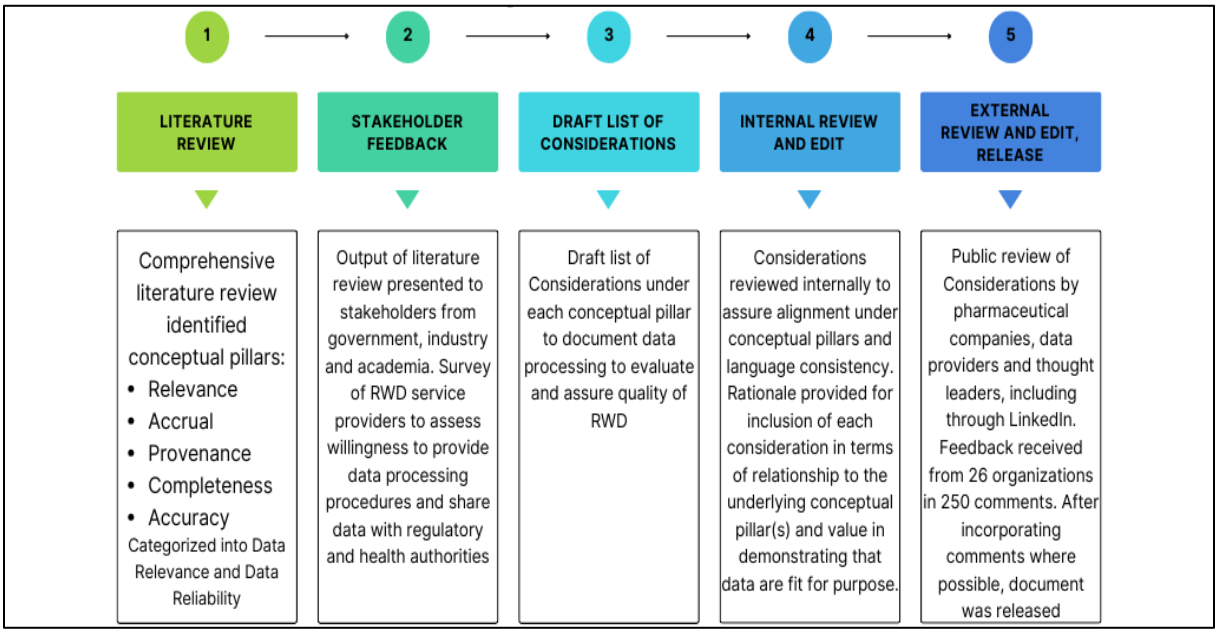

Figure 1: Audit Readiness Considerations Developmental Process - Adapted from TransCelerate (TransCelerate 2023)

The conceptual pillars identified included:

1) Relevance of the data captured for elements needed to address the specific research question,
2) Accrual process by which the data are captured and patients are included in the specific study along with operational definitions of key variables,
3) Provenance, including origins of data elements and processing from data collection to analytic file
4) Completeness of key data elements to address the research question
5) Accuracy and consistency in storage of the data fields.

The conceptual pillars 3 through fall under the category of data quality and reliability. Documentation for these categories would be referred to in protocols and clinical study reports, and detailed in statistical analysis plans (SAPs), data management plans, standard operating procedures (SOPs) and data dictionaries (TransCelerate Biopharma, 2023).

Our working group of leaders from the International Society of Pharmaco-epidemiology (ISPE), International Society for Pharmaceutical Outcomes Research (ISPOR) and TransCelerate developed an editable form as a framework to ensure comprehensive documentation based on the TransCelerate Considerations for Audit Readiness and extensive team experience. (TransCelerate Biopharma, 2023), FDA guidance documents (FDA 2018, 2021, 2023; EMA 2023a), the EMA data quality

framework (EMA 2023b) This framework acts as an inventory for submitters and a rubric for reviewers.

3. Results

Following detailed discussions on the contents of a data package submission, we composed a framework (outlined in Appendix I: Contents for RWE Data Package Submission) to ensure that a regulatory submission package addresses aspects outlined in the regulatory guidance documents. Below we briefly discuss the contents of each section.

A. Relevancy And Fit-For-Purpose (Appendix I: Section I)

Before considering relevancy and fit-for-purpose aspects, it is critical to understand the scientific research question and regulatory context (decision). The PICOTS framework (Population, Intervention, Comparator, Outcomes, Time, and Setting) can be used to clearly articulate a research question and identify the critical study elements (Richardson et al., 1995; Ritchey & Girman, 2020). Other frameworks for developing well-articulated research questions and protocols for RWE studies have been published (Gatto et al., 2019; Gatto et al., 2022; Wang et al., 2023). Sponsors should include justification of how the data source adequately captures the target population that would use the medicinal product and how any biases in population identification are mitigated. The setting (e.g., inpatient, outpatient) strongly influences whether a population is representative. The treatment groups (intervention, comparator) and outcome(s) must be adequately captured and relevant to the scientific research question and regulatory decision.

B. Feasibility of RWD Source in General (Appendix I: Section II)

Documentation should be provided for each data source assessed for feasibility of generating RWE (FDA, 2021). A data source should have detailed documentation of internal data processing and data quality procedures for a regulatory reviewer to adequately evaluate whether the data source is fit-for-purpose to address the scientific question and regulatory decision. (FDA, 2021; Ritchey & Girman. 2020).

Specifically, documentation related to the following should be included:

a. the type of data (claims, electronic medical records, registry or other) and coding system(s) used
b. how data is checked for errors during aggregation and extraction, transform and loading (ETL)
c. how any linkages of patient information are performed across healthcare (e.g., primary vs. specialist care) over time and verified, and
d. how analytic datasets are created from source data. (FDA, 2018)

This documentation can be included in statistical analysis plans (SAPs), data management plans, data dictionaries, specific SOPs, and /or output from quality control reports. Additional SOPs and/or other source documents should clearly describe quality assurance procedures at each step in the data source lifecycle. Finally, the ability to generate audit trails in data processing from data collection to creation of analytic datasets is an important consideration.

C. Feasibility for a Specific Study (Appendix I: Section III)

The feasibility of a data source for a specific study is contextual, depending on whether

study elements can be adequately identified with good data quality in the context of the scientific research question. (Ritchey & Girman 2020). When evaluating a data sources, the following are helpful tools to use to define key study design elements:

- Structured Pre-approval and Post-approval Comparative study design framework to generate valid and transparent real-world evidence (SPACE) ; and
- Structured Process for Identifying Fit-For-Purpose Data (SPIFD) as well as PICOTS and START-RWE checklists (Gatto et al., 2019; Gatto et al., 2022; Wang et al., 2023).

For study-specific RWD source feasibility, specific considerations should include study population selection, proper application of inclusion/exclusion criteria, validity of the outcome and exposure as defined in the data sources as well as the breadth and scope of relevant covariate captures (FDA, 2021; Ritchey & Girman, 2020). Sufficient accuracy of each of the key study elements should be justified in the data package.

Study Population

Alignment between the study population and target population and the risk of systematic differences that could be introduced during the selection (and result in bias) should be addressed. In addition, to selecting the right individuals, their time and data in the study from the period in their life, patient journey, and care settings relevant to the research question should be captured. The method of identification of study subjects or patient records based on a specific exposure, disease, characteristics, or a combination of such variables should be reported in the protocol/SAP and data package, including any applicable accuracy metrics.

Exposures, Study Outcomes and Other Study Variables

The database should include variables capturing the underlying condition and its severity, and any variables needed to characterize the population and exposures accurately, including inclusion/exclusion criteria, outcomes, patient and disease characteristics, comorbidities, facility and provider characteristics, and potential confounding factors or modifiers. Exposures should be reported with sufficient detail, including capture and accuracy of timing(s) of exposure(s), duration(s) and dose(s), if relevant. Outcomes should be reported along with validation methods and accuracy metrics, with high specificity and high positive predictive value especially important for rare diseases or safety studies (Appendix I: VI) (Weinstein et al., 2023). Timing of outcomes should be captured over a sufficient duration to address study follow-up, be accurate enough for study purposes and be described in study documentation.

D. Missing Data (Appendix I: Section IV)

Considerations around missing RWD differ dramatically from clinical trials. RWD gaps can occur when (1) data are not present at the source (e.g., a clinician does not ask a question or document a data point), (2) data becomes missing during data transformation, or during conversion to a common data model, and (3) variables are created in the analytic file based on raw data that contains missing data.

To the extent known, ETL practices should detail and document both expectedly and/or systematically missing (e.g., age >80 years masked for patient privacy) and unexpectedly missing data along with any approaches taken to understand the impact of missing data on results (sensitivity analyses, imputation methods, censoring). Methods should be consistent with whether data are assumed missing at random (MAR)

or missing not at random (MNAR). Steps to mitigate the risk of missing data should be addressed.

A key limitation of insurance claims is missing data when patients move in and out of health plans. The same is true for facility-based claims or EMR, when patients seek care out of network. Both instances can lead to bias in study data. Data are strongly influenced by reimbursement policies, health system policies and patient reporting, which are rarely documented within a database. Assumptions about non-present codes should be explicitly stated in the SAP. Derived measures (i.e. values calculated for analysis or stemming from multiple different "source" values) should be checked for coding errors or inclusion of missing data in calculations. This is particularly important in settings where analytic datasets are assembled from multiple aggregated source databases. Mechanisms of gaps (if any), and steps to evaluate the impact of these gaps on results should be described.

Other causes of RWD gaps fall under governance relating to privacy considerations and de-identification (Appendix I: VIII), proprietary practices inhibiting full transparency of RWD transformation, or contractual agreements or costs that limit the scope or completeness of data for end users. Limited or missing documentation describing data transformation inhibits complete understanding of how ETL processes may have influenced results.

E. Handling of Unstructured Data (Appendix I: Section V)

In EHR, about 80-85% of data is stored in unstructured format and may include medical imaging, radiology reports, clinical notes or discharge summaries (Kong, 2019). These data are often critical when defining study outcomes. If unstructured data were used in studies supporting regulatory submissions, the below should be documented:

- All variables generated from clinicians' notes, or electronic medical records should be listed, with primary/secondary outcomes and key covariates/stratification variables highlighted
- Variables derived based on the vendor's standard offering should be differentiated from those based on a study-specific abstraction process
- Methodologies used to extract information from unstructured sources (e.g. natural language processing (NLP) or artificial intelligence (AI) algorithm) should be clearly delineated
- If unstructured data was manually abstracted or reviewed via chart review or adjudication for the purposes of validation, SOPs or charters should clearly define the processes and accuracy metrics.

Reference to assumptions and verification approaches should be provided in the vendor's SOPs and/or Data Management Plans, along with where to find information on specific variable algorithms.

F. Validation (Appendix I: Section VI)

Operational definitions that are used to select the population (i.e., inclusion/ exclusion criteria), exposures (e.g., treatment groups), patient outcomes and confounding variables should be provided explicitly in the protocol or SAP. If the coding algorithm is not clearly written and previously validated in the literature - ideally in a similar data source - it may be necessary to justify the accuracy and completeness of the operational definition. (Weinstein et al., 2023). Any study-specific or external validation method along with

accuracy metrics should be described. For computer algorithms, the detailed operational process needs be described (supervised vs. unsupervised algorithm), in addition to data sources used. Any impacts on data quality and steps undertaken to prevent bias should also be discussed.

The operational definition used to select and define patients must be appropriately sensitive and must avoid erroneously identifying those without the condition or exposure (appropriate specificity), while recognizing that no algorithm has perfect sensitivity and specificity. If literature or well-tested definitions are not available, a validation study should be conducted to assess these properties (with report). Otherwise, the Sponsor would be expected to justify that the operational definitions applied in the RWD source are adequate for the purpose intended by the research questions. Sensitivity analyses along with quantitative bias analyses can help justify that results are unchanged when operational definitions are varied.

G. Data Provenance (Appendix I: Section VII and end III)

For studies conducted in data sources from a specific data source provider (DSP), transparency and provenance of the data from data collection to creation of an analytic dataset are of critical importance. Many RWD datasets pass through three layers: 1) data collection and transformation at point of care, 2) data aggregation, ETL and de-identification at a data vendor or curator, and 3) data acquisition and ETL at the research organization (Sponsor). Data provenance is the ability to clearly trace back a given variable or data element in an analytic dataset to the source of collection (e.g. point of care) through each layer of data collection, handling, and processing. Transparency is important for credibility in the methodologies to process and validate healthcare data and to help identify potential sources of bias that may be introduced. SOPs for each step of data processing are essential.

If a DSP utilizes third party vendors to (a) collect/aggregate data from healthcare providers or (b) conduct cleaning/managing of data, additional SOPs may be needed related to vendor management and oversight. Additionally, audit trails capturing transformations or manipulations of source data elements in creating analytic datasets are important for understanding data provenance. Audit trails should include a description of source data, data manipulations or computationally handling in cloud computing environments, and meta-data related to output. Furthermore, codebooks or data dictionaries of meta-data are critical and should be available in submissions to obtain an understanding of the genesis of the analytic dataset.

H. De-Identification and Patient Privacy (Appendix I: Section VIII)

A fundamental concept of data provenance relates to the desire by regulators and auditors to trace data back to the original source (i.e., data verification). Similarly, the “Accrual” pillar that requires data privacy and patient consent is paramount for any prospective data collected as part of a primary data collection study. Permanently de-identified data of DSPs pose little concern to subject privacy but prohibits tracing data back to the original source. Additionally, patient consent for tracing data to its source was not obtained nor is a patient identifiable to obtain it. In these cases, DSP documentation and SOPs for receiving, aggregating and curating the data and the de-identification processes must be

made available. The Sponsor can provide traceability from the time of receipt of de-identified data from DSP to the analytic file. In addition, the de-identified file received by the sponsor/researcher and the analytic file could be made available to FDA, if the sponsor has obtained agreement for that from the DSP.

In RWE studies where the study or part of the study requires data collection, informed consent may be required and there is a need to balance data provenance with the obligation to maintain subject privacy. As an example, where limited data are easily anonymized, increased quantities of collected data inextricably increases the risk for possible re-identification of subjects, especially in the case of rare diseases. The prevalence of AI computing tools further increases the possibility of re-identification. These risks require careful evaluation and possible implementation of risk mitigation techniques. The approach and considerations should be clearly outlined within the data package submission, including a summary of risks to both data provenance and data privacy, and applicable risk mitigation steps.

No universally accepted definition exists for "de-identification" and "anonymization" with de-identification only vaguely described within the US Privacy Rule (45 CFR § 164.514 (b)) and the EU General Data Protection Regulation (https://gdpr.eu/tag/gdpr/). These terms are sometimes used interchangeably within published research; other manuscripts imply varying degrees of subject confidentiality (Chevrier et al., 2019). In instances where the terms are contrasted, de-identification generally refers to the removal or pseudonymization (replacing data that might identify a patient with similar, but not the exact same data) of multiple types of data elements and is less likely than anonymized data to be re-identified. In the least stringent of definitions, anonymized data requires only the removal of the subject's name whereas the most stringent requires removal of all data that would potentially link the subject to the data contained within their medical records (Chevrier et al., 2019).

While de-identification and anonymization are addressed conceptually within the TransCelerate Audit Readiness document (TransCelerate BioPharma, 2023), definitions for neither are established. The intended definitions should be clearly stated within the data submission package, along with methods and controls to prevent re-identification of subjects.

3. Discussion

While the FDA and EMA have used real-world evidence on safety of products for decades, evaluations of effectiveness in real-world data have only rarely been the basis for approval or labelling changes. The 21st Century Cures Act required the FDA to consider use of real-world evidence, resulting in a series of guidance documents issued from the Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER) and the Center for Devices and Radiological Health (CDRH). While these guidance documents greatly help elucidate FDA thinking on RWD, the level of detail and specificity in terms of what is needed for submissions remains vague. Further, based on regulatory drug and biologics approvals in the last five years, RWE is still relatively infrequently used for decision-making. Therefore, little precedence exists to help sponsors understand requirements (Beaver et al., 2018; Agrawal et al., 2023; Jahanshahi et al., 2021; Purpura et al., 2022; Sola-Morales et al., 2023; Izem et al., 2022).

TransCelerate's initiative makes significant headway in further characterizing what the FDA needs in order to evaluate the data quality and relevance of RWD for regulatory purposes. The audit readiness considerations outlined issues that should be addressed in submission of RWD packages to FDA when using RWE for critical regulatory decisions. However, the level of specificity and detail was still unclear. We outlined a dynamic framework that would give more guidance on expectations for sponsors and DSPs, when they are developing a data package on the relevancy and reliability of RWD for a regulatory decision and specific research question.

The itemized form (Appendix I) could be useful not only as a guide to the contents of needed documentation in a filing, but also in briefing packages and discussions with FDA about the use of RWE to support specific regulatory decisions. The form can be used for FDA submissions but could be useful for other regulatory submissions if the documentation needed is deemed comparable. While the focus is on regulators, our team believed that if that high bar were met, the contents would likely also be acceptable to Health Technology Assessment (HTA) bodies. Again, this is a tool to ensure appropriate documentation is included in RWE submissions, not a replacement for such documentation.

This dynamic framework would change based on the specific context of the submission, including the specific research question, the data elements needed to address the research question and the regulatory decision. In addition, there may need to be greater scrutiny of data quality and relevance when the treatment effects are expected to be small to moderate as the signal to noise ratio will be lower. Likewise, the data elements needed for a study may drive the specific issues to be addressed. For example, if death is an outcome and claims data are being used, linkage to the national death index and/or state obituary data may be needed to fully capture death and especially cause of death in an outpatient setting.

Berger et al. (2024) published a tool that had general similarities to the TransCelerate effort, addressing broad categories of RWD authenticity, transparency, relevance, accuracy and track record. Likewise, a systematic literature review described attributes of data quality framework with software implementations in R (Schmidt et al., 2021) that were similar to the EMA (EMA, 2023b). Others have documented attributes and components of data quality and released code in R to perform the assessments (Blacketer et al., 2021; Liaw et al., 2021). While useful as frameworks, these tools have focused on data quality assessments on the data, and not what specific documentation and SOPs might be needed for a regulatory submission on RWE.

Our framework is based substantially on the TransCelerate work, which in turn built on the draft guidance RWE documents issued by the U.S. Food & Drug Administration between 2018 and 2023. At the end of 2023, a draft guidance on use of RWE for regulatory decisions on medical devices was released and covers the concepts from the 2017 prior guidance documents in more detail (CDRH & CBER, 2023). Specifically, the draft guidance indicates that RWD are fit-for-purpose for informing a specific regulatory decision if both relevancy and reliability can be demonstrated. The data source should include the data elements needed to address the research question and regulatory decision, including the intervention (e.g.,

medical device), the population and outcomes of interest and key covariates, and there should be sufficient longitudinal data to inform the regulatory decision. The data should be timely and generalizable to the population who will use the device for treatment. The guidance went into detail about the documentation on data processes and procedures to ensure accuracy, consistency, and completeness of data that are required for review of RWE regulatory submissions to FDA. In addition, documentation on the operational definitions of all variables included in the analysis or used to define data elements (population, device, and outcome) are needed. Our framework is highly consistent with this new draft guidance (CDRH & CBER, 2023), despite it being released, after the TransCelerate Considerations were developed (TransCelerate BioPharma, 2023).

The framework we present is not prescriptive but provides clarity on the potential documentation that could be helpful to health authorities reviewing RWE submissions for decision-making. The specific details of the framework would change based on the data issues for a specific study and regulatory decision. For example, documentation may differ substantially for a natural history study that used primary data collection, compared to a study based on claims data. This could help sponsors who are creating documentation for submission to regulatory authorities, to guide discussions between sponsors and regulatory authorities about the inclusion of RWE in a submission. Additionally, the framework could clarify the different types of information data providers must include with their data to assist in the submission of RWE data. Eventually, artificial intelligence (AI) will be able to generate all the available documentation for data quality, accuracy, completeness and relevance. However, until data providers and industry generate such documentation for specific databases, human oversight will be needed to ensure the documentation is correct and sufficient.

4. Conclusion

Real world data (RWD) is increasingly used to support regulatory decisions, but what is needed to justify data quality and relevance for the regulatory decision and research question has been unclear. Using the considerations for audit readiness released by TransCelerate in 2023, we developed guidance for developing a regulatory data package for justifying RWD relevance and quality for implementing a real-world evidence protocol. The data package should delineate the research question and regulatory decision, describe each RWD source evaluated for feasibility, and the type of RWD, its accuracy, completeness, and adequacy in capturing study variables, the details of the data processing and the extraction, transformation and loading (ETL) documentation, as well as audit trails. The dynamic framework should: a) help data providers understand what their customers need for RWE submissions, b) assist researchers in understanding what documentation to gather to prepare for regulatory RWE submissions, c) guide sponsors in preparing submissions packages using RWD, and d) aid regulators in their review of RWE submissions by identifying documentation sponsors should provide. This tool may not be sufficient for all purposes, for all regulators, or even for all RWE submissions, but should serve as a guide or a starting point to help with the organization of documentation needed.

5. Acknowledgements

The authors thank Rachele Hendricks Sturrup of Duke-Margolis Center for Health Policy for her helpful comments and Andre Araujo for his initial efforts in getting this work implemented.

6. Author Contributions

Manuscript:
Cynthia J Girman, Christina Mack, Rachel Leahy, Dana Teltsch, Morgan A Marks, Grazyna Lieberman, Molly L. Aldridge, and Christopher Atzinger.
Research Design & Implementation:
Cynthia J Girman, Christina Mack, Rachel Leahy, Dana Teltsch, Morgan A Marks, Grazyna Lieberman, and Christopher Atzinger.
Documentation Analysis: Cynthia J Girman, Christina Mack, Rachel Leahy, Dana Teltsch, Morgan A Marks, Grazyna Lieberman and Christopher Atzinger.

7. Previous Presentation

Portions of this work were presented in poster format (selected as spotlight poster) at the 2024 Annual Meeting of the International Society for Pharmaco-epidemiology in Berlin, Germany, August 24-28, 2024, and also as a poster at 2024 Annual meeting of International Society for Pharmaceutical Outcomes Research in Atlanta, GA, May 5-8, 2024

8. Funding

No funding was received for this work.

9. Conflict of Interest

Cynthia J Girman is founder of CERobs Consulting, which provides services to the pharmaceutical and medical device industries. Christina Mack is an employee of IQVIA, which provides services to healthcare entities. Rachel Leahy is principal of Paradigm Bioscience and works with CERobs Consulting, both of which provides regulatory compliance and quality services to the pharmaceutical industry. Dana Teltsch is a full-time employee of Takeda and the outgoing chair of the Digital Epidemiology Special Interest Group of ISPE. Morgan A Marks is an employee and current stockholder of Merck and Co, and owns stock in Moderna Therapeutics and Pfizer. Grazyna Lieberman was employed by Genentech while contributing to the TransCelerate efforts, previously employed by N-Power Medicine, and received meeting support from Genentech. Grazyna Lieberman holds stock in Roche. Christopher Atzinger is an employee of Astellas.

10. References

Agrawal, S., Arora, S., Amiri-Kordestani, L., de Claro, R. A., Fashoyin-Aje, L., Gormley, N., Kim, T., Lemery, S., Mehta, G. U., Scott, E. C., Singh, H., Tang, S., Theoret, M. R., Pazdur, R., Kluetz, P. G., & Beaver, J. A. (2023). Use of Single-Arm Trials for US Food and Drug Administration Drug Approval in Oncology, 2002-2021. JAMA oncology, 9(2), 266–272. https://doi.org/10.1001/jamaoncol.2022.5985

Beaver, J. A., Howie, L. J., Pelosof, L., Kim, T., Liu, J., Goldberg, K. B., Sridhara, R., Blumenthal, G. M., Farrell, A. T., Keegan, P., Pazdur, R., & Kluetz, P. G. (2018). A 25-Year Experience of US Food and Drug Administration Accelerated Approval of Malignant Hematology and Oncology Drugs and Biologics: A Review. JAMA oncology, 4(6), 849–856. https://doi.org/10.1001/jamaoncol.2017.5618

Berger, M. L., Crown, W. H., Li, J. Z., & Zou, K. H. (2024). ATRAcTR (Authentic Transparent Relevant Accurate Track-Record): a screening tool to assess the potential for real-world data sources to support creation of credible real-world evidence for regulatory decision-making. Health Services and Outcomes Research Methodology, 24(3), 348-365.

Berger, M.L., Daniel, G., Frank, K., Hernandez, A., McClellan, M. (2017). A Framework for Regulatory Use of Real-World Evidence. Duke Margolis Institute for Health Policy. https://healthpolicy.duke.edu/publications/framework-regulatory-use-real-world-evidence.

Blacketer, C., Defalco, F. J., Ryan, P. B., & Rijnbeek, P. R. (2021). Increasing trust in real-world evidence through evaluation of observational data quality. Journal of the American Medical Informatics Association: JAMIA, 28(10), 2251–2257. https://doi.org/10.1093/jamia/ocab132

Center for Devices and Radiological Health and Center for Biologics Evaluation and Research. (2023). Use of real world evidence to support regulatory decision making for medical devices: Final guidance for industry and Food & Drug Administration staff U.S. Food & Drug Administration. U.S. Department of Health and Human Services Food & Drug Administration. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/draft-use-real-world-evidence-support-regulatory-decision-making-medical-devices

Chevrier, R., Foufi, V., Gaudet-Blavignac, C., Robert, A., & Lovis, C. (2019). Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review. Journal of medical Internet research, 21(5), e13484. https://doi.org/10.2196/13484

European Medicines Agency (EMA). (2023a). Real-world evidence framework to support EU regulatory decision-making Report on the experience gained with regulator-led studies from September 2021 to February 2023. EMA/289699/2023. https://www.ema.europa.eu/system/files/documents/report/real-world-evidence-framework-support-eu-regulatory-decision-making-report-experience-gained_en.pdf

European Medicines Agency (EMA). (2023b). Data quality framework for EU medicines regulation. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/data-quality-framework-eu-medicines-regulation_en.pdf

Gatto, N. M., Campbell, U. B., Rubinstein, E., Jaksa, A., Mattox, P., Mo, J., & Reynolds, R. F. (2022). The Structured Process to Identify Fit-For-Purpose Data: A Data Feasibility Assessment Framework. Clinical pharmacology and therapeutics, 111(1), 122–134. https://doi.org/10.1002/cpt.2466

Gatto, N. M., Reynolds, R. F., & Campbell, U. B. (2019). A Structured Preapproval and Post approval Comparative Study Design Framework to Generate Valid and Transparent Real-World Evidence for Regulatory Decisions. Clinical pharmacology and therapeutics, 106(1), 103–115.

https://doi.org/10.1002/cpt.1480

Izem, R., Buenconsejo, J., Davi, R., Luan, J. J., Tracy, L., & Gamalo, M. (2022). Real-World Data as External Controls: Practical Experience from Notable Marketing Applications of New Therapies. Therapeutic innovation & regulatory science, 56(5), 704–716. https://doi.org/10.1007/s43441-022-00413-0

Jahanshahi, M., Gregg, K., Davis, G., Ndu, A., Miller, V., Vockley, J., Ollivier, C., Franolic, T., & Sakai, S. (2021). The Use of External Controls in FDA Regulatory Decision Making. Therapeutic innovation & regulatory science, 55(5), 1019–1035. https://doi.org/10.1007/s43441-021-00302-y

Kennedy-Martin, T., Curtis, S., Faries, D., Robinson, S., & Johnston, J. (2015). A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. Trials, 16, 495. https://doi.org/10.1186/s13063-015-1023-4

Kong H. J. (2019). Managing Unstructured Big Data in Healthcare System. Healthcare informatics research, 25(1), 1–2. https://doi.org/10.4258/hir.2019.25.1.1

Liaw, S. T., Guo, J. G. N., Ansari, S., Jonnagaddala, J., Godinho, M. A., Borelli, A. J., de Lusignan, S., Capurro, D., Liyanage, H., Bhattal, N., Bennett, V., Chan, J., & Kahn, M. G. (2021). Quality assessment of real-world data repositories across the data life cycle: A literature review. Journal of the American Medical Informatics Association: JAMIA, 28(7), 1591–1599. https://doi.org/10.1093/jamia/ocaa340

Purpura, C. A., Garry, E. M., Honig, N., Case, A., & Rassen, J. A. (2022). The Role of Real-World Evidence in FDA-Approved New Drug and Biologics License Applications. Clinical pharmacology and therapeutics, 111(1), 135–144. https://doi.org/10.1002/cpt.2474

Richardson, W. S., Wilson, M. C., Nishikawa, J., & Hayward, R. S. (1995). The well-built clinical question: a key to evidence-based decisions. ACP journal club, 123(3), A12–A13.

Ritchey, M. E., & Girman, C. J. (2020). Evaluating the Feasibility of Electronic Health Records and Claims Data Sources for Specific Research Purposes. Therapeutic innovation & regulatory science, 54(6), 1296–1302. https://doi.org/10.1007/s43441-020-00139-x

Rodriguez, F., Califf, R. M., & Harrington, R. A. (2019). Consequences of Slow Progress Toward Pragmatism in Randomized Clinical Trials: It Is Time to Get Practical. JAMA cardiology, 4(11), 1129–1130. https://doi.org/10.1001/jamacardio.2019.3922

Schmidt, C. O., Struckmann, S., Enzenbach, C., Reineke, A., Stausberg, J., Damerow, S., Huebner, M., Schmidt, B., Sauerbrei, W., & Richter, A. (2021). Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. BMC medical research methodology, 21(1), 63. https://doi.org/10.1186/s12874-021-

01252-7

Sola-Morales, O., Curtis, L. H., Heidt, J., Walsh, L., Casso, D., Oliveria, S., Saunders-Hastings, P., Song, Y., Mercado, T., Zusterzeel, R., Mastey, V., Harnett, J., & Quek, R. G. W. (2023). Effectively Leveraging RWD for External Controls: A Systematic Literature Review of Regulatory and HTA Decisions. Clinical pharmacology and therapeutics, 114(2), 325–355. https://doi.org/10.1002/cpt.2914

TransCelerate BioPharma. (2023). Assuring audit and inspection readiness – Considerations for the use of RWD and RWE in regulatory decision-making. https://www.transceleratebiopharmainc.com/wpcontent/uploads/2023/12/Assuring-Audit-and-Inspection-Readiness-Considerations-for-the-use-of-RWD-and-RWE-in-Regulatory-Decision-Making_12.11.23.pdf

U.S. Food & Drug Administration. (2018). Framework for FDAs Real World Evidence Program. https://www.fda.gov/media/120060/download

U.S. Food & Drug Administration. (2021). Real-world data: Assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products: Guidance for Industry. https://www.fda.gov/media/152503/download

U.S. Food & Drug Administration. (2023). Considerations for the use of real-world evidence to support regulatory decision-making for drug and biological products: Guidance for Industry. https://www.fda.gov/media/171667/download

Wang, S. V., Pottegård, A., Crown, W., Arlett, P., Ashcroft, D. M., Benchimol, E. I., Berger, M. L., Crane, G., Goettsch, W., Hua, W., Kabadi, S., Kern, D. M., Kurz, X., Langan, S., Nonaka, T., Orsini, L., Perez-Gutthann, S., Pinheiro, S., Pratt, N., Schneeweiss, S., Williams, R. J. (2023). Harmonized Protocol Template to Enhance Reproducibility of hypothesis evaluating real-world evidence studies on treatment effects: A good practices report of a joint ISPE/ISPOR task force. Pharmaco-epidemiology and drug safety, 32(1), 44–55. https://doi.org/10.1002/pds.5507

Weinstein, E. J., Ritchey, M. E., & Lo Re, V., 3rd (2023). Core concepts in pharmacoepidemiology: Validation of health outcomes of interest within real-world healthcare databases. Pharmaco-epidemiology and drug safety, 32(1), 1–8. https://doi.org/10.1002/pds.5537

Appendix I: Contents for RWE Data Package Submission

<table>
<tr><th colspan="4">I. Relevancy and Fit-For-Purpose</th></tr>
<tr><td colspan="2">Research Question:
To assess < safety/effectiveness > of < intervention >
compared to < control, if any > in patients with < condition >
Population: < condition > Intervention:
Comparator: Outcome(s):
Setting (in-/out-patient):
Duration and study period:</td><td>Regulatory Context:
Is study supporting:
□ Product approval
□ New indication
□ Other labeling expansion
□ Natural history</td><td>Outcomes are related to:
□ General effectiveness/ safety
□ Safety
□ Adherence
□ Treatment patterns</td></tr>
<tr><th colspan="4">II. Feasibility of Data Sources in general + link to Feasibility Report [Add columns as needed for additional data sources]</th></tr>
<tr><td colspan="2">Data Source: _____</td><td colspan="2">Data Source: _____</td></tr>
<tr><td>□ EHR
□ Registry</td><td>□ Claims
□ Other: _____</td><td>□ HER
□ Registry</td><td>□ Claims
□ Other: _____</td></tr>
<tr><td colspan="2">□ Single health system □ Multiple health systems</td><td colspan="2">□ Single health system □ Multiple health systems</td></tr>
<tr><td colspan="2">Intent of data collection:
☐ Patient care ☐ Billing ☐ Research study
Coding system (e.g. ICD-10):
Data captures indicated population:</td><td colspan="2">Intent of data collection:
☐ Patient care ☐ Billing ☐ Research study
Coding system (e.g. ICD-10):
Data captures indicated population:</td></tr>
<tr><td colspan="2">Documentation provided for:
□ General accuracy
□ Completeness
□ Out-of-network care capture
□ Consistency over time
Total database # patients:
Common date model available?
If yes, which: _____</td><td colspan="2">Documentation provided for:
□ General accuracy
□ Completeness
□ Out-of-network care capture
□ Consistency over time
Total database # patients:
Common date model available?
If yes, which: _____</td></tr>
<tr><td colspan="2">Documentation (if available):
□ ETL procedures SOP
□ Error/Range checking & correction
□ Plausibility checks
□ Deduplication
□ De-identification SOP</td><td colspan="2">Documentation (if available):
□ ETL procedures SOP
□ Error/Range checking & correction
□ Plausibility checks
□ Deduplication
□ De-identification SOP</td></tr>
<tr><td colspan="4">*-Continue to next page-*</td></tr>
</table>

Appendix I: Contents for RWE Data Package Submission (Continued)

III. Feasibility for Specific Study + link to Feasibility Report *[Add columns as needed for additional data sources]*	
Data Source: _____	Data Source: _____
Inclusion/exclusion definable? *Structured codified data exists for:* □ Population □ Appropriate age/gender? □ Outcome □ Exposure *Documentation contains key study elements (e.g., population, intervention/comparator, outcome, exposures, covariates, timeframe, key confounders) to determine:* □ *Accuracy* □ *Completeness* *Total study population (N): _____* *N with outcome (blinded): _____* *N with treatment: _____* *Average follow-up (median): _____* *Study period: _____* *Continuous enrollment during lookback?* *Changes in study period†: _____*	Inclusion/exclusion definable? *Structured codified data exists for:* □ Population □ Appropriate age/gender? □ Outcome □ Exposure *Documentation contains key study elements (e.g., population, intervention/comparator, outcome, exposures, covariates, timeframe, key confounders) to determine:* □ *Accuracy* □ *Completeness* *Total study population (N): _____* *N with outcome (blinded): _____* *N with treatment: _____* *Average follow-up (median): _____* *Study period: _____* *Continuous enrollment during lookback?* *Changes in study period†: _____*
-Continue to next page-	

Appendix I: Contents for RWE Data Package Submission

STUDY LEVEL	
Frequency of refresh: _____ Last date of refresh: _____ Most recent data: _____	Frequency of refresh: _____ Last date of refresh: _____ Most recent data: _____
Trail to analytic file available? Is linkage‡ applied? If so, what source: _____ For which variables: _____ Data dictionary utilized*: _____ *At Sponsor's Receipt:* □ ETL Procedure SOP □ Data integrity control- SOP □ De-Duplication □ Checks converting to CDM □ Conformance to CDM □ Plausibility & consistency checks □ If so, describe: _____ □ Range/error checks □ If used, describe: _____ *Analytic File:* □ Outlier/range checks □ Corrections documented □ Transformation defined □ Viewable audit trail + analytic file □ Audit trail (ALCOA + at minimum) □ Special handling of labs □ Availability of program code QC □ Checks performed on code □ Data is sharable with the FDA	Trail to analytic file available? Is linkage‡ applied? If so, what source: _____ For which variables: _____ Data dictionary utilized*: _____ *At Sponsor's Receipt:* □ ETL Procedure SOP □ Data integrity control- SOP □ De-Duplication □ Checks converting to CDM □ Conformance to CDM □ Plausibility & consistency checks □ If so, describe: _____ □ Range/error checks □ If used, describe: _____ *Analytic File:* □ Outlier/range checks □ Corrections documented □ Transformation defined □ Viewable audit trail + analytic file □ Audit trail (ALCOA + at minimum) □ Special handling of labs □ Availability of program code QC □ Checks performed on code □ Data is sharable with the FDA
† Changes in Medical Practice/Diagnosis/Treatment, Coding Changes/Updates, Coverage of Products, Prior Authorizations/Tier ‡ If linkage performed, include how (Deterministic, Probabilistic) and Procedures Used to check matching accuracy in DMP; if more than one data source is used, documentation should be provided for the linked dataset. * Details of data coding and data dictionary should be included in DMP	
-Continue to next page-	

Appendix I: Contents for RWE Data Package Submission (Continued)

IV. Missing Data	
Data Provider Aggregation: *Confirm completeness of data documented for:* □ Primary Care Visits □ Specialist Visits Hospitalizations □ Out-of-Network Care □ Specific outcomes of interest and death □ Exposure(s)- treatment and comparator if relevant	**Handling of Missing Data in Analysis:** *Are elements of algorithms missing to define:* □ Population □ Outcome □ Exposures If imputation performed, what type: Is data missing at random?

V. Handling of Unstructured Data

Is information from clinicians' notes in analysis? - If yes, variables generated using the vendor's standard offering:
Variables generated using a study-specific abstraction process:
What approaches of the vendor's standard offering were used for this study? (Check all that apply):
□ Human abstraction □ Technology assisted human abstraction □ NLP □ Machine Learning
Are vendor's SOPs/DM plans available for review/audit? -
Were study-specific abstraction processes described in the protocol or in separate study document? Is a report summarizing the implementation of the process included in a separate report? -
What approaches were used in the study-specific abstraction process? (Check all that apply):
□ Human abstraction □ Technology assisted human abstraction □ NLP □ Machine Learning

VI. Validation

Justify adequate measurement properties of:
- □ Algorithm/coding for outcomes *(Link separate validation report or publication if applicable.)* Link:
- □ Algorithm/coding for exposure *(Link separate validation report or publication if applicable.)* Link:
- □ Algorithm/coding for key confounders *(Link separate validation report or publication if applicable.)* Link:

Was validation performed or published: _____ If so, complete the following: _____

Metric	Study measures (% [95% CI])				
	Primary outcome: _____	Primary Exposure: _____	Key Confounder 1: _____	Key Confounder 2: _____	Key Confounder 3: _____
PPV					
Sensitivity					
Specificity					

PPV: Positive predictive value

VII. Provenance	
Data custodians over lifecycle of data documented? Source data manipulations described? Controls/processes ensuring data integrity given?	Audit trail processes described? Changes? Audit trails/metadata complete? Fulfill ALCOA+ Audit trails/metadata viewable/accessible? Can analytic files be shared with regulators?

Appendix I: Contents for RWE Data Package Submission (Continued)

VIII. Risk of De-identification	
How risk was mitigated at Data Provider level: Registries (if applicable; state whether public or private):	How risk was mitigated at analysis stage (e.g., cells n<5 not reported):