{2024-AUGUST-30}

RWD Guidelines for Programming and Analysis Processes

Revision History

Version	Date	Summary
V1.0	30Aug2024	Initial version

Table of Contents

1. Introduction

As per FDA[1.1], 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. Examples include data derived from Electronic Health Records (EHRs), Claims and/or billing data, Product and/or disease registry data, other data sources that can inform on health status (e.g., data collected from wearables, patient-generated data).

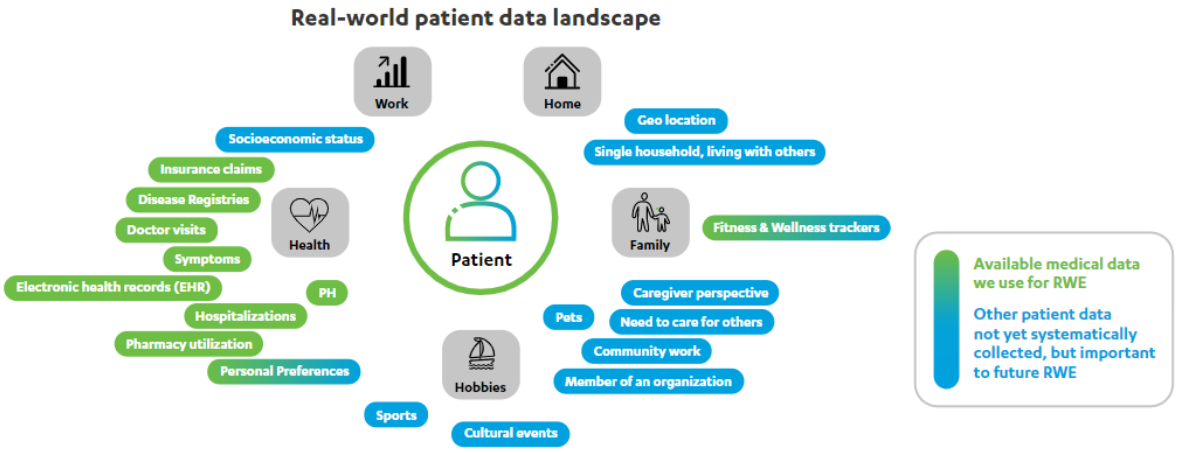


Figure 1.1: Real World patient data landscape. Source: Janssen RWD Playbook, Janssen R&D [not publicly available]

Real-World Evidence (RWE) as defined by FDA, is the clinical evidence regarding the usage and potential benefits, or risks of a medical product derived from analysis of RWD.

Similar definitions have been used by other regulatory bodies, such as EMA.

1.1. Real World Data and Evidence

Randomized Controlled Trials (RCTs) are considered the gold-standard for drug approvals and label claims, however, the demand for RWE is on the rise. Factors contributing to the increasing importance of RWE include but are not limited to[1.2]:

- New regulatory initiatives
- Access to patient data through medical record databases and disease registries
- Increased interest in:
  - Patient-specific benefits,
  - Providing cost-effective large-scale means of monitoring effectiveness and safety (where randomized trials may not be feasible).
- Bridge the evidence gap between clinical research and practice.

RWD have been used successfully in the following cases where RCTs are challenging:

- Rare disease
- Pediatric studies
- Randomization not possible due to toxicity (oncology)
- Ethical issues with continuing placebo (e.g., Covid-19 vaccine)

Global regulators, such as the US Food & Drug Administration (FDA), Japan's Pharmaceutical and Medical Devices Agency (PMDA), the European Medicines Agency (EMA), the UK's Medicines and Healthcare products Regulatory Agency (MHRA) and China's National Medical Products Agency (NMPA) are increasingly interested in leveraging the potential of RWD to complement TCTs with RWE to support regulatory decision making across the product lifecycle[1.3]. The FDA is accepting observational data to support efficacy determinations and have already issued approvals of new indications for approved drugs[1.4]. EMA is assessing the use of registry data[1.5] and other RWD to support various pre and post market authorization activities shown in Figure 1.2 below.

As the healthcare landscape continues to evolve, it is imperative that organizations evolve with it and various functions within R&D act proactively to provide the best solutions for supporting RWE activities in line with regulatory expectations. Functions like Data Management, Programming and Statistics play a key role in an organization's ability to leverage and continue to build on cross-functional RWE capabilities and competencies, integrating valuable learnings, new perspectives, and best practices

across the RWE portfolio. This paper is intended to act as a guide to the Statistical Programming community in achieving these objectives in a pragmatic, efficient and productive manner and importantly towards the end goal of serving patients in bringing new treatments to the market.

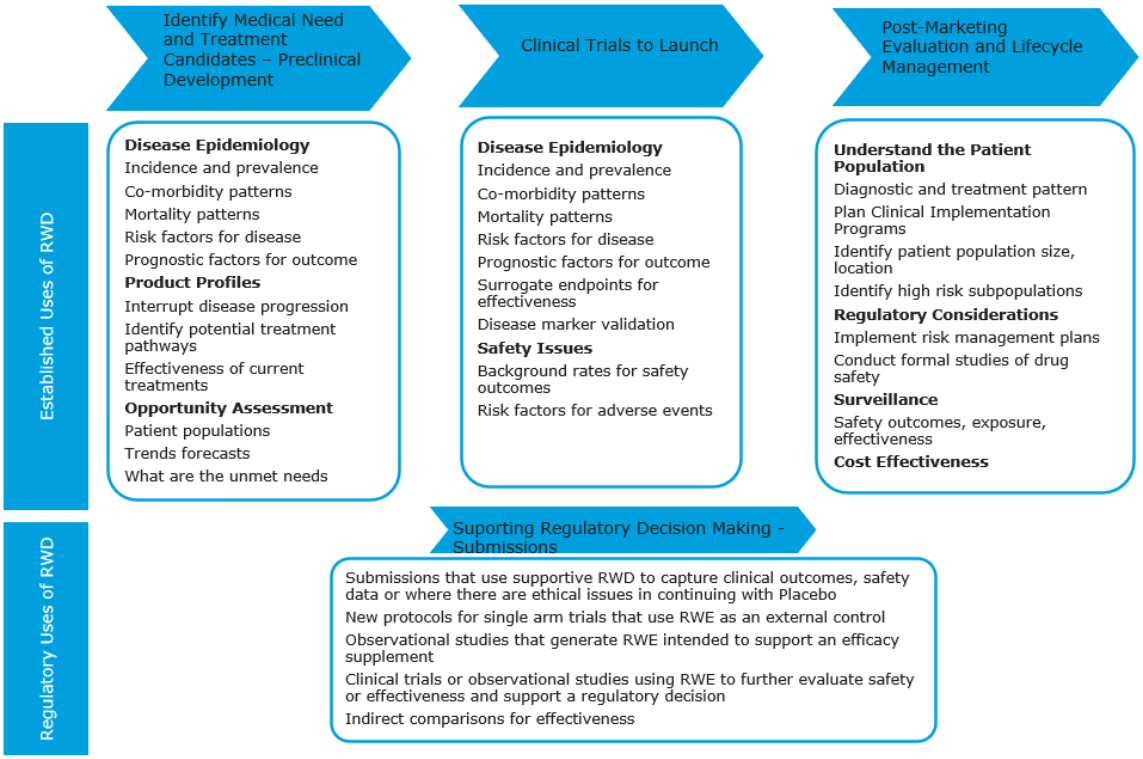


Figure 1.2: Practical uses of RWD during drug development through post-marketing. Source: Janssen RWD Playbook, Janssen R&D [not publicly available]

1.2. Purpose of Real-World Data

Randomized Controlled Trials (RCTs) are considered the gold standard in clinical research for evaluating the efficacy and safety of interventions. One key strength of RCTs is randomization, which helps minimize selection bias, ensuring that both known and unknown confounding variables are evenly distributed between groups, thereby strengthening the internal validity of the study. By controlling for confounders through randomization, RCTs provide a strong foundation for making causal inferences

about the effects of interventions, which makes them essential for informing clinical practice guidelines, regulatory decisions, and healthcare policies.

Despite their strengths, RCTs also have limitations, such as high costs, ethical considerations, and potential challenges in generalizing findings to broader populations. There are a few advantages for integrating RWD with RCT:

Broader patient presentation: RCTs often have strict eligibility criteria, limiting the diversity and representation of patients, hence may not fully capture how a drug performs in diverse patient groups or under real-world conditions, potentially leading to limited applicability of findings. By integrating RWD, which includes data from real-world clinical practice, a broader range of patients with varying demographics, comorbidities, and disease severity can be included in analyses. This enhances the generalizability of findings to real-world populations.

Long term safety and effectiveness: RCTs are typically conducted over a limited timeframe with a focus on short-term outcomes, hence relying solely on RCTs may not provide sufficient data on rare adverse events or long-term outcomes. RWD, especially from longitudinal studies or post-marketing surveillance, provides valuable insights into the long-term safety and effectiveness of drugs in routine clinical practice. This helps in understanding the drug's performance over extended periods and in real-world settings.

Comprehensive evidence generation: Integrating RWD with RCTs allows for a more comprehensive evaluation of a drug's efficacy, safety, and real-world impact. RWD can complement RCT data by providing additional evidence on patient outcomes, treatment patterns, adherence, and healthcare resource utilization.

Cost effectiveness and efficiency: Leveraging existing RWD sources, such as electronic health records and claim databases, alongside RCTs can optimize resource allocation and in the long term reduce the cost and time required for drug development. RWD can facilitate post-market research and support label expansions by providing supplementary evidence.

External Controlled Arm: Despite RCTs having obvious appeal in clinical research, it has some well-known limitations. In rare diseases or in diseases where no effective standard-of-care treatments are available, it is not often feasible or ethical to recruit patients to control groups. An uncontrolled, single arm trial where all participants receive the investigational treatment is more appealing in these

scenarios. However, without an internal control group, assessments are performed by making indirect comparisons, which may be suboptimal.
Use of a comparator based on data collected outside of a study, referred to as an external control or synthetic control group, could offer a compromise between uncontrolled trials and RCTs in certain context[1.6]. An external control group could consist of patients treated at an earlier time (sometimes referred to as an historical control) or patients treated during the same period of time but in a different setting (sometimes referred to as a contemporaneous control).

In summary, combining RWD with RCTs in drug development enhances patient representation, provides insights into long-term outcomes and real-world impact, supports comprehensive evidence generation and improves cost-effectiveness. This integrated approach can lead to more informed and robust evaluations of drug efficacy and safety, benefiting patients, healthcare providers, and stakeholders involved in drug development and healthcare delivery.

The purposes of the use of RWD as part of a regulatory submission / decision may include the following:

- To provide evidence in support of effectiveness or safety for a new product approval
- To provide evidence in of support of labeling changes for an approved drug, including:
  - Add or modify an indication
  - Change in dose, dose regimen, or route of administration
  - Use in a new population
  - Add comparative effectiveness information
  - Add safety information
  - Other labeling change

- To be used as part of a post-marketing requirement to support a regulatory decision
  - Post approval Safety Study
  - Post approval Effectiveness Study

1.3. Scope

This White Paper is primarily intended to guide statistical programmers, especially those who are new to

RWD/RWE by providing recommended best practices and/or minimum requirements for supporting RWE activities that may lead to regulatory submission. Although the primary audience is the statistical programmers, other functions that are typically associated with RCTs may benefit also. In general, "non-regulatory submission" activities involving the use of RWD are not in scope when it comes to applying the full set of requirements as recommended by Regulatory Health Authorities (RHAs). This may include:

- exploratory analyses (for internal decision making / publications)
- support with analyses for manuscripts and publications
- Market Access-Health Technology Assessments (HTA) submissions

However, if Statistical Programming or other functions are involved in the data analysis activities for the above-mentioned considerations, the best practices from this white paper are still relevant.

This paper chiefly discusses the guidance published by US FDA for submission of RWD to support marketing application of new drug or biological products or new indication for drugs already approved. However, the topics discussed in this paper are equally relevant for other regulatory health authorities too when it comes to submission of RWD/RWE for marketing applications.

This paper is organized in a way to mimic the processes that are typical of a RCT submission to RHAs. This way the reader will have a general understanding of the processes related to RWE studies in comparison to processes relevant to RCTs.

1.4. Types of RWE Studies

The term “RWD” or “RWE” encompasses a broad range of study design, providing more specifics about data sources and nature of the study.

RWD sources (e.g., registries, EHRs, administrative and medical claims databases) can be used for data collection and, in certain cases, to develop analysis infrastructure to support many types of study designs to develop RWE, including, but not limited to, randomized trials (e.g., large sample trials, pragmatic clinical trials) and observational studies (prospective or retrospective).

1.4.1. Observational studies

Observational studies are non-interventional clinical study designs whereby patients receive treatment during routine medical care that are not guided by research protocols, even if laboratory or imaging procedures are done per protocol.

A non-interventional study is a type of study in which patients received the marketed drug of interest during routine medical practice and are not assigned to an intervention according to a protocol. Examples of non-interventional study designs include:

- Observational **cohort studies**, in which patients are identified as belonging to a study group according to the drug or drugs received or not received during routine medical practice, and subsequent biomedical or health outcomes are identified
- **Case-control studies**, in which patients are identified as belonging to a study group based on having or not having a health-related biomedical or behavioral outcome, and antecedent treatments received are identified.

A **retrospective cohort study** is a study which identifies the population and determines the exposure/treatment from historical data (i.e., data generated prior to the initiation of the study and therefore after the outcome events have occurred). The variables and outcomes of interest are determined at the time the study is designed.

This type of study includes database research, review of records or analysis of electronic healthcare records where all the events of interest have already happened, and making secondary use of existing data which were not collected for the study purposes.

This type of study can stand alone or be combined with prospective data capture, creating hybrid designs that can be more time and cost-efficient than complete de novo data capture.

In a **prospective cohort study**, the population of interest is identified at the start of the study, patients are enrolled prior to the occurrence of outcome events and followed prospectively over time. The start of the study is defined as the time at which the research protocol for the specific study question is initiated.

A **registry**, is defined as a prospective, non-interventional organized collection of human data within a particular disease, group or other "at risk" special patient population (e.g., cancer, pregnancy, organ transplant) with design characteristics as follows:

- A systematic collection of defined events or exposures

- Is conducted in a defined population in one or more specific geographic areas
- The data collected could be either for a defined period of time or indefinitely.

The questions typically addressed in registries range from purely descriptive questions aimed at understanding the characteristics of patients who develop the disease and how the disease generally progresses, to highly focused questions intended to support decision making. Registries focused on determining clinical effectiveness or cost-effectiveness or assessing safety or harm are generally hypothesis driven and concentrate on evaluating the effects of specific treatments on patient outcomes.

A **registry-based study** is an investigation of a research question using the infrastructure of (a) new or (an) existing registry(-ies) for patient recruitment and data collection. A registry-based study may be a clinical trial, or a non-interventional study.
A registry-based study may apply primary data collection and/or secondary use of data collected in a patient registry for another purpose than the given study.

Rare diseases represent a highly heterogeneous group of disorders with high phenotypic and genotypic diversity within individual conditions. Due to the small numbers of people affected, there are unique challenges in understanding rare diseases and drug development for these conditions, including patient identification and recruitment, trial design, and costs. Natural history data and RWD play significant roles in defining and characterizing disease progression, final patient populations, novel biomarkers, genetic relationships, and treatment effects. A **natural history study** is a preplanned observational study intended to track the course of the disease. Its purpose is to identify demographic, genetic, environmental, and other variables (e.g., treatment modalities, concomitant medications) that correlate with the disease's development and outcomes. Natural history studies are likely to include patients receiving the current standard of care and/or emergent care, which may alter some manifestations of the disease. Disease registries are a frequent platform to acquire data for natural history studies. For rare diseases, natural history studies play an important role in identifying appropriate patient populations and clinical outcome assessments and biomarkers, and in the design of externally controlled studies. Beyond their role in drug development, natural history studies may also benefit patients with rare diseases by establishing communication pathways, identifying disease-specific centers of excellence, facilitating the understanding and evaluation of current standard-of-care practices, evaluating signs and symptoms of a disease to improve diagnosis, and identifying ways to improve patient care. Patients included in natural history studies may sometimes be used as historical controls for studies that lack an internal control, thus allowing the effectiveness of the study treatment to be determined.

1.4.2. Trials in Clinical Practice Settings

Pragmatic clinical trials (PCTs), sometimes called practical clinical trials, are designed to evaluate the comparative effectiveness of interventions within routine clinical settings. These trials are "pragmatic" because they focus on understanding how interventions work in real life as opposed to "explanatory" where the goal is to determine if and how an intervention works. Key aspects of PCTs are broad population inclusion, study design and data collection procedures that minimally disrupt routine clinical care encounters, and an emphasis on patient-centered health outcomes. Pragmatic Randomized Trials (PRTs) represent a hybrid between traditional randomized controlled clinical trials (that are the gold standard for regulatory decision making), and pragmatic, observational research studies that are often used to generate real-world evidence. A well-designed PRT that maximizes external validity, but also controls for confounding (including but not limited to selection bias) in order to maintain high levels of internal validity, could theoretically be used to generate evidence that could meet regulatory requirements. Evidence from these trials is specifically relevant when treatment options do already exist for the disease under study and when the real-life situation, including extraneous factors, is expected to influence the treatment effect.

The process of randomization in Randomized Controlled Trials (RCTs) removes confounding by known and unknown factors. However, in the case that a randomized control arm is not possible, an **External Controlled Arm (ECA)** may be an option for estimating comparative treatment effect.

Typically, the external control arm uses data from past traditional clinical trials, but in some cases, RWD have been used as the basis for external controls. Using external controls has limitations, including difficulties in reliably selecting a comparable population because of potential changes in medical practice, lack of standardized diagnostic criteria or equivalent outcome measures, and variability in follow-up procedures. Collection of RWD on patients currently receiving other treatments, together with statistical methods, such as propensity scoring, could improve the quality of the external control data that are used when randomization may not be feasible or ethical, provided there is adequate detail to capture relevant covariates[1,7].

Hybrid Prospective Designs (e.g., concurrently randomized control as well as external control) allow the integration of a traditional randomized controlled trial with pragmatic design aspects to collect real-world

data on patients. This design preserves the benefit of randomization, provides real-world outcome data while potentially accelerating product development and lowering the cost of data collection and patient follow-up.

2. Planning and Study Set up

The objective of conducting clinical studies of a drug is to distinguish the effect of the drug from other influences, such as spontaneous change in the course of the disease, placebo effect, or biased observation. When relying on a non-interventional study (e.g., EHR data generated during routine clinical care analyzed using a cohort study design), the inference(s) drawn may be incorrect if based on estimates that are affected by (1) confounding (e.g., due to noncomparable treatment groups) or (2) other forms of bias.

Accordingly, before choosing a non-interventional study design for a study intended to support regulatory decisions regarding the safety and effectiveness of a product, sponsors and researchers should consider how likely it is that such a study design and its conduct will be able to distinguish a true treatment effect from other influences.

Sponsors should identify and address commonly encountered challenges when considering the use of a non-interventional study for regulatory decision-making[2.1].

Although Statistical programmers are generally not involved in the early stages of a clinical study, it's still recommended to have an awareness of the processes upstream especially if regulatory submission is planned.[2.1]

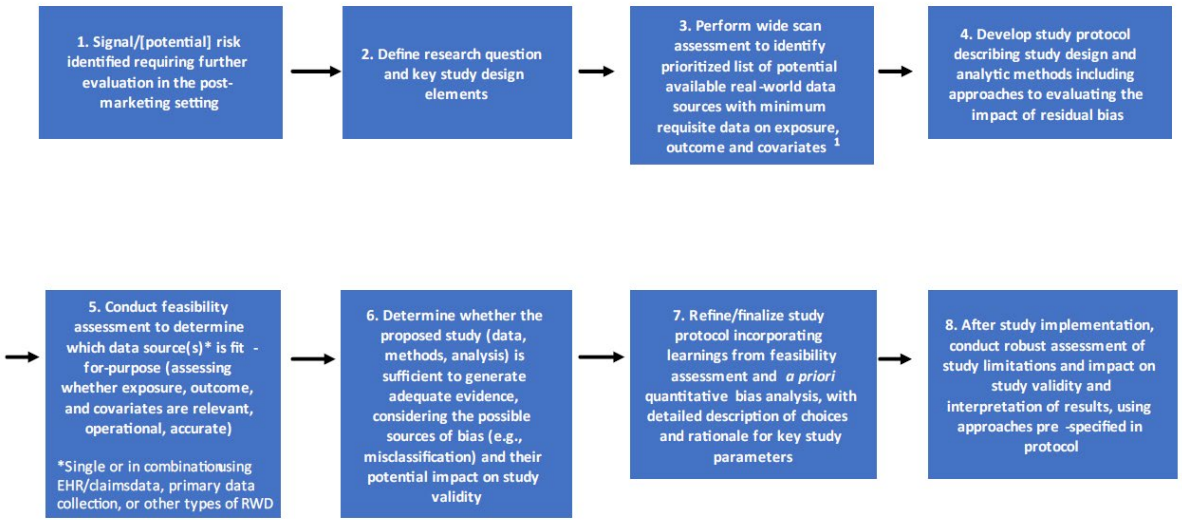


Figure 2.1: A framework for developing adequate evidence using fit for purpose real world data to address regulatory questions on drug safety. Gatto NM et al. The Structured Process to Identify Fit-for-Purpose Data: A Data Feasibility Assessment Framework.

2.1. Framing of the Research Question

The research question is a concise statement of the study purpose and the prespecified hypotheses to be tested; the purpose of the study may also be to generate hypotheses for future research. It is generally assumed that the research question has been defined and agreed ahead of Statistical Programming engagement. In the protocol, researchers should document decisions about the study design and the types of data required/available. Careful formulation of the research question will highlight unknowns that will need to be addressed through information derived from the feasibility assessment and this information may further refine the question and drive protocol development.[2.2]

2.2. Feasibility Assessment

A feasibility assessment is a systematic process to identify fit-for-purpose data to address a specific research question. When conducting a study-level feasibility assessment, a key goal is to describe and compare the reliability and relevance of the data sources assessed for the research question. Statistical Programmers may be involved in this assessment.

Feasibility assessments should be structured in at least two phases:

- an initial scan to determine whether the available data sources will suffice and to narrow down data source options, and
- a subsequent, more comprehensive feasibility assessment of the narrowed-down data sources.

In the early stages of designing a non-interventional study, sponsors should discuss with the regulators the expectations regarding access to patient level or analytic data sets. Sponsors should obtain any required agreements relevant patient-level/analytic data that will be required for submission by the regulator.

Submission of the feasibility assessment report can either be a standalone document, an annex to the protocol, or used as context for design decisions in the protocol. The final approach should comply with applicable regulatory requirements. Detailed frameworks, templates, and checklists for conducting feasibility assessments are available in scientific publications.[2,3]

2.3. Engagement with Regulatory Health Authorities

Considering the evolving and diverse regulatory frameworks, early engagement with regulatory agencies is highly recommended. Prior to conducting non-interventional studies, sponsors are advised (but this is fast becoming a requirement) to submit the draft Protocol and Statistical Analysis Plan (SAP) to the relevant RHA. For FDA, sponsors should finalize the study protocol, including the research question of interest and rationale for the study design, before initiating study conduct.

FDA strongly encourages sponsors to engage with the Agency in the early stages of designing a non-interventional study and to provide sufficient information needed to clarify expectations related to the design and proposed conduct of their study. Although detailed information on every attribute described may not be available or feasible to include at the time of early engagement with FDA, successful proposals for non-interventional study designs should satisfactorily address key attributes such as, summary of the proposed approach, study design, data sources, and analytic approach, as applicable.

When the available data sources do not support proposals that can satisfactorily address each of these attributes, alternative study designs should be considered.

2.4. Protocol Development and Review

Development of the protocol (and SAP) occurs with alignment with the research question. The feasibility analysis will guide the development of the protocol and facilitate the discussion with Regulatory Health Authorities, HTA bodies and other parties. A major factor in bolstering confidence in RWE studies and ultimately producing RWE strong enough for decision making by regulators is the selection of fit-for purpose data prior to protocol finalization.

The protocol should contain detailed descriptions on study design, data sources and data types, target population, exposure, outcomes and covariates and the proposed analytical approaches. The study protocol and SAP should specify the data provenance (curation and transformation procedures used throughout the data life cycle) and describe how these procedures could affect data integrity and the overall validity of the study.

Although Statistical Programmers are not part of the Protocol Review Committee (PRC), a Statistical Programming representative needs to be involved in the review of Protocols describing RWE studies that are planned to be submitted. Statistical Programmers should focus their review on areas that could impact the ability to interpret, transform, analyze, or pool data.

2.5. Real World Data Checklist

At the study start up, Statistical Programmers who are involved in the study can produce a checklist for internal use. Although the primary user and owner of this checklist is anticipated to be the statistical programming group at sponsor and CRO organizations, other groups such as biostatistics, data management, project management, and RWE group may benefit from the content of this checklist.

The validity of the content of this checklist can be revisited after the clinical study reaches milestones such as finalization of protocol/SAP and other study specific relevant milestones.[2.4]

2.5.1. Purpose of RWD Checklist

Since the use of RWD is increasing rapidly in both interventional as well as observational clinical studies, it is recommended that users apply this checklist in studies that include elements of RWD/RWE. Any such use of RWD in a clinical study can have implications in terms of how such data can be submitted to regulatory bodies. Please refer to Appendix 1.1 and example cases in Appendix 1.2 for more details about how the checklist can be used. Keeping these aspects in mind, as well as considering its inclusion in broader RWE guidance documentation, this checklist serves the following purposes:

1) It captures initial information pertaining to the nature of real- world evidence study, providing helpful assistance and insights to statistical analysis and reporting functions. The options selected in this checklist are expected to help the statistical programming group get a thorough view of nature of the study in reference to use of Real-World Data.

2) It helps the user understand real world elements of the data and submission related aspects depending on factors like sources of data and efforts involved in meeting regulatory submissions requirements.

3) It serves as a preliminary guiding document to initiate dialogue with other stakeholders within the organization towards study planning, statistical analysis, and reporting activities.

4) The outcome of the checklist can help the user with decision making for RWD submission planning. Some examples of this are explained as a part of case examples in Appendix 1.2.

3. RWD Study Council

As clinical research continues to add RWD to study designs more and more, the evolution of multistakeholder teams must take place as well. The existing RCT clinical team is no longer sufficient to properly navigate RWD considerations, so an equivalent is needed. A proposed name for this new team is a RWD Study Council. Programmers will find themselves donning various titles depending on the organization structure, but there are general principles that can be highlighted as RWD Study Councils go from sparse to common in the coming years.

3.1. Purpose

This chapter will unpack how a RWD Study Council functions, including general team member roles and how a responsibility assignment matrix can assist in visibility across the team, with an emphasis on details relevant to programmers. Therefore, given the scope of this White Paper for programmers, other roles will need to expand beyond what is presented here, as it is not comprehensive or exhaustive for every role on the RWD Study Council.

3.2. RACI Matrix

One of the most common responsibility assignment matrices is the RACI matrix, which will be utilized here as a recommended practice for task clarity. There are numerous sources describing the RACI acronym, so one has been selected from a Forbes Advisor article and shared below:

"**Responsible** designates the task as assigned directly to this person (or group of people). The responsible person is the one who does the work to complete the task or create the deliverable. Every task should have at least one responsible person and could have several."

"The **accountable** person in the RACI equation delegates and reviews the work involved in a project. Their job is to make sure the responsible person or team knows the expectations of the project and completes work on time. Every task should have only one accountable person and no more."

"**Consulted** people provide input and feedback on the work being done in a project. They have a stake in the outcomes of a project because it could affect their current or future work."

"**Informed** folks need to be looped into the progress of a project but not consulted or overwhelmed with the details of every task. They need to know what's going on because it could affect their work, but they're not decision makers in the process."

The RACI matrix is a proven resource and will be especially helpful as RWD Study Councils get up and running. A general RACI matrix template is proposed below, including common contributors/roles and activities, followed by company-specific examples.

3.3 Contributors/ Roles

The chart in Figure 3.1 shows possible role dependencies and the roles in the top level are included in Table 3.1.

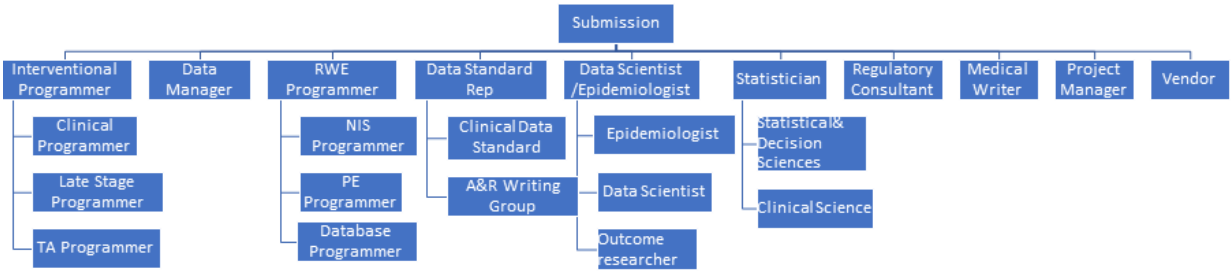


Figure 3.1. RWD - Hierarchy of Study Roles

TA-Therapeutic Area

PE- PharmarcoEpidemiology

NIS- Non Interventional Study

A&R- Analysis Report

The roles for participants in RWD are included as separate columns in Table 3.1. Each company may use different naming conventions based on the department structure and related internal processes. It is important to note that both RCT programmer and RWE programmer are required to support RWE studies. The individual rows present the main tasks that can be considered for the submission preparation process.

There are four main study phases listed as: Study Initiation (including kick-off meetings, project management tasks, etc.), Study Execution (including the data management and data handling rules/standards, etc.), Submission Preparation (including the different regulatory agency requirements, etc.), and last phase is study close-out (including final submission packages and sharing data).

3.4. General RACI Table

Study Phase	Activity/Roles	Interventional Programmer	Data Manager	RWE Programmer	Data Standard Rep	Data Scientist /Epidemiologist	Statistician	Regulatory Consultant	Medical Writer	Project Manager	Vendor Partner
Study Initiation	**Planning**										
	Complete Roles and Responsibilities (RACI)										
	Project Management/ Kick-off Meeting										
	Identifying the RWD sources										
	Feasibility Assessments										
Study Execution	**Data Handling**										
	Data Transfer/Data Share Agreements										
	Considerations										
	Data Standardization Plan										
	Coding Versions										
	Milestone dates agreement										
	Data Analytics										
	SDTMs/ADaMs and Metadata										
	Tables/Listings/Figures										
	SDTMs/ADaMs										
Submission Preparation	**Submission Package Review**										
	Submission Requirements										
	Regulatory Reviews										
Close-out	**Final Submission Package**										
	Sign off final documents										
	Submit to regulatory agency										

Table 3.1. RACI- RWD - Study Roles and Activities

Table 3.1 serves as a template to construct a RACI matrix. It provides a high level of categories for individual contributors and activities involved in RWE submission. Below are two examples to show different approaches to adapting a RACI template based on company's requirements and/or organizational structure.

3.4.1. RACI Table Examples

Company A

There is no separation between RWD and RCT programmers in company A. Clinical Programming (CP) and Statistical Programming (SP) constitute the programming group which deals with both types. In case of fully RWE studies, the Data Sciences (DS) group are the driver for analysis (equivalent to stats in RCT) and they come to CP and SP when submission is planned.

Activity										
	CDS&T	DM	RM-CM	CP	SP	RMW	SDS	DS/Epi	CS	GTL/ GPL
Identifying the RWD sources	I	I		I	I		R, A	R, A	I	I
Feasibility Assessment: mini analysis plan	I	I	I	I	C	C	C, R	R, A	I	
Feasibility Assessment: Determining Data Relevancy		I	I	I	C	C	C, R	R, A	I	I
Feasibility Assessment: Data Reliability analysis		R	I	R	R		R, A	R, A	I	I
Feasibility Assessment and Report		C	C	C	C, R	R	R, A	R, A	I	I
Protocol (using appropriate template)	C	C	C	C	C	R, A	C	C	C	I
Vendor Engagement	C	C, R	I	C	C	I	C	C, R	I	R
Data Ingestion Source Data Flows		R, A		C	C		I	I	I	
Coding	C	R, A		C	C		I	I	I	
Database Releases		R, A		C	C		C	C	C	
Creation of Mapping File		I		R, A	C		C	C	I	
Creation of SDTMs or CDM	C	R, A		R, A	C		I	I	I	
Creation of SAP		I		C	C		R, A	I	I	
Creation of DPS	C	I		I	R, A	C	R	C, I	C	
Creation of ADaMs	C	I		I	R, A		C	I	I	
Data linkage from more than one Databases	C	R		C	R, A	I	R	R	R	
Data Integrations		I		I	R, A		C	C	C	
Standards used	C	I		R, A	R, A		R, A	R, A	I	
Determination of waiver requirement for SDTMs	C	I		R, A	R		C	I	I	
Determination of waiver requirement for ADaMs		I		I	R, A		C	I	I	
Study Data Standardization Plan	R	C		C	C					
Privacy / Legal Considerations	C	I		I	I		R, I	R, I	R, I	R
Data De-identification	R									

Abbreviations: CDS&T = Clinical Data Standards & Transparency, DM = Data Management, CP = Clinical Programming, SP = Statistical Programming, RMW = Regulatory Medical Writing, SDS = Statistics and Decision Sciences, DS = Data Sciences, Epi = Epidemiology, CS = Clinical Sciences.

Company B

There is separation between RWD and RCT programmers in company B. RWD Programmers –fall under two different roles, Therapeutic Area (TA) and Regulatory RWE Programming (RRP). TA programmers are given an SAP to program against. They are doing the actual regulatory study. RRP is there to guide the TA programmers through the data standards, train on the templates, excel spec, P21, reviewer's guides, and show how to use the SAS macros to make sure they are fulfilling regulatory and company B requirements and processes.

Activity								
	TA Programming Lead	TA Programmer(s)	TA Epidemiology Lead	RRP Lead	RRP Programmer(s)	PMO	Cross Functional Partners: RCT Programmer(s)	Consultants: SDTM / ADaM Standards
Study Initiation								
Submit RRP Parent DAC Request for Regulatory Study	I		R	I		I		
Submit PMO Request	I		I	R		I		
Schedule PMO Kick-off Meeting	I	I	I	I	I	R		
Study Setup								
Create MS Teams for Regulatory Study	C	I	I	C	I	R	I	
Create Smartsheet Risk Register	I		I	I		R		
Create Project Status Dashboard Slide	I		I	I		R		
Create Question and Decision Log	I		I	I		R		
Assess Data Submission Needs (e.g. SDTM, ADaM, P21E license, etc)	R	I	C	R	I		C	C
Complete Roles and Responsibilities Form	R	I	R	C	I			
Complete Data and Folder Access Form	R	I	I	C	I			
Submit RWD Engineering Child DAC Request	R			I		I		
Raw Data Access and Preparation	I	I	I	I	I	I		
Create Regulatory Project Folder	I	I	I	I	I			
Populate Regulatory Project Folder (templates, etc)	I	I	I	A	R			
Assess Clean Room Needs	R		R	R				
Open and Setup Clean Room	I	I	I	I	I			
Submit Approved Clean Room Requests	C	I	R	C	I			
Fulfill Clean Room Requests	A	R	C	C	C			
Study Request Review								
Study Analysis								
Ad Hoc Analysis								
Study Sign-off								

Abbreviations: TA = Therapeutic Area, RRP = Regulatory RWE Programming, PMO = Project Management, RCT = Randomized Controlled Trial, DAC = Data Analytic Centre.

3.4.2. Conclusion

As the need for RWD Study Councils continues to grow, clarity is needed for team members to understand roles and responsibilities. The RACI matrix provides this clarity and can be adapted to various organizational structures as needed. RWD programmers can offer this solution to their study teams if it does not already exist.

4. Data Provenance, Ethics and Data Privacy

4.1 Data Provenance

Data provenance refers to the complete documentation of a dataset's lineage—including its origin, collection context, transformation processes, and any subsequent modifications. As a form of metadata, provenance serves as an audit trail that enables verification of a dataset's authenticity, supports reproducibility, and ensures transparency and accountability in scientific research (Buneman et al., 2001). Provenance plays a critical role in validating the integrity of clinical and real-world data and is essential for evidence generation in regulatory and health technology assessments. In parallel, data privacy refers to the protection of sensitive personal information from unauthorized access, use, or disclosure. In today's data-driven landscape—particularly in the context of clinical trials, digital health, and genomics—data privacy is increasingly challenged by the need for openness, traceability, and interoperability.

While data provenance enhances transparency and trust, it can also pose privacy risks, especially when provenance metadata includes sensitive contextual details that may inadvertently reveal personal or institutional identities (Simmhan et al., 2005). Thus, there is a critical need to balance transparency with confidentiality. Provenance frameworks must be designed to respect data minimization principles and incorporate privacy-preserving techniques, such as de-identification, anonymization, tokenization, hashing, and encryption, to ensure that sensitive information is not exposed (El Emam & Arbuckle, 2013).This balancing act becomes particularly important in clinical research, where enabling data sharing and leveraging emerging technologies such as artificial intelligence, distributed learning, and cloud-based analytics introduces new risks—including data theft, re-identification, and unauthorized manipulation. De-identification remains the most common privacy-enhancing method, yet its effectiveness varies across regulatory contexts. For instance, GDPR requires that anonymized data be irreversibly de-linked from individuals, while other frameworks, such as the U.S. HIPAA Safe Harbor standard, apply more pragmatic thresholds based on re-identification risk (Malin et al., 2013; Tene & Polonetsky, 2013).

4.2 Data Ethics

Data ethics broadly refers to the moral principles guiding how individuals, organizations, and institutions collect, manage, protect, and use data. It encompasses core values such as fairness, transparency, accountability, and respect for privacy, emphasizing that data governance should serve both individual

rights and the public good (Floridi & Taddeo, 2016). Ethical data practices are particularly critical in healthcare, where the misuse or overuse of sensitive health information can lead to discrimination, stigmatization, and erosion of public trust.

4.3. Data Privacy

4.3.1 Adhering to Data Privacy Regulations

Ensuring compliance with data privacy regulations is essential for safeguarding individual data, fostering stakeholder trust, and mitigating legal and financial risks. Key legislative frameworks—such as the General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), and Health Insurance Portability and Accountability Act (HIPAA)—establish stringent requirements for data collection, processing, and disclosure. Adherence to these regulations necessitates a comprehensive approach that includes respecting data subject rights (e.g., access, rectification, and erasure), conducting Data Protection Impact Assessments (DPIAs), implementing timely breach notification protocols, and ensuring accountability across third-party data processing arrangements.[4.1 4.2 4.3 4.4 4.5]

- General Data Protection Regulation (GDPR) (2018) - A comprehensive data protection law enacted by the European Union, GDPR governs the collection, processing, and storage of personal data. It strengthens individual privacy rights, mandates lawful data handling practices, and applies extraterritorially to any entity processing EU citizens' data.[4.6]
- Personal Information Protection and Electronic Documents Act (PIPEDA) (2000) - Canada's federal privacy law for private-sector organizations, PIPEDA governs how businesses collect, use, and disclose personal information during commercial activities. It emphasizes individual consent, data access rights, and the requirement to safeguard personal information.[4.7]
- The Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) -A foundational U.S. regulation that protects individuals' medical records and other personal health information (PHI). The HIPAA Privacy Rule sets national standards for the use and disclosure of PHI by covered entities and gives patients' rights over their health data, including rights to access, amend, and request restrictions on its use.[4.9]

Growing global emphasis on transparency has led regulatory agencies to implement policies that govern the public disclosure of clinical trial data. The following key initiatives illustrate how major regulators have shaped the data sharing landscape:

- EMA Policy 0043 (2010) and Policy 0070 (2015)

The European Medicines Agency (EMA) introduced Policy 0043 to enable public access to documents held by the agency, including regulatory submissions. Policy 0070, implemented in 2015, advanced transparency by proactively publishing clinical study reports submitted in marketing authorization applications, while protecting personal and commercially confidential information.[4.11]

- Health Canada Public Release of Clinical Information (2019)
  Implemented in 2019, Health Canada's PRCI initiative mandates the proactive publication of anonymized clinical information from drug and medical device submissions following final regulatory decisions. This policy aims to enhance transparency, support independent research, and align with international best practices, while ensuring the protection of personal and confidential business information.[4.8]
- EU Clinical Trial Regulation (2022)
  Effective from January 31, 2022, the EU Clinical Trial Regulation harmonizes the assessment and supervision processes for clinical trials across EU Member States. It introduces the Clinical Trials Information System (CTIS), a centralized portal facilitating single-submission applications for multinational trials, thereby enhancing efficiency, transparency, and participant safety.[4.10]

4.3.2. Statistical Disclosure methods

Statistical disclosure control (SDC) encompasses a set of techniques designed to minimize the risk of re-identifying individuals from statistical outputs while preserving the analytical value of the released data. These methods are particularly critical in the dissemination of aggregate statistics, microdata, and synthetic datasets derived from sensitive sources, such as clinical trials, electronic health records, or administrative databases (Hundepool et al., 2012).

At its core, SDC seeks to uphold data confidentiality and ensure that information disseminated to researchers, policymakers, or the public complies with ethical principles and legal frameworks such as the General Data Protection Regulation (GDPR), HIPAA, and national statistical legislation. SDC techniques are used not only to prevent direct identification but also to mitigate the risks of inferential disclosure, where individuals can be indirectly identified through combinations of attributes or auxiliary information (Willenborg & de Waal, 2001).

Types of Statistical Disclosure Control Techniques

SDC methods can be broadly categorized into two types:

Input SDC (applied to microdata): These methods are used before statistical analysis and include:

- Data swapping
- Top- and bottom-coding

- Noise addition
- Data suppression
- Record aggregation or micro aggregation

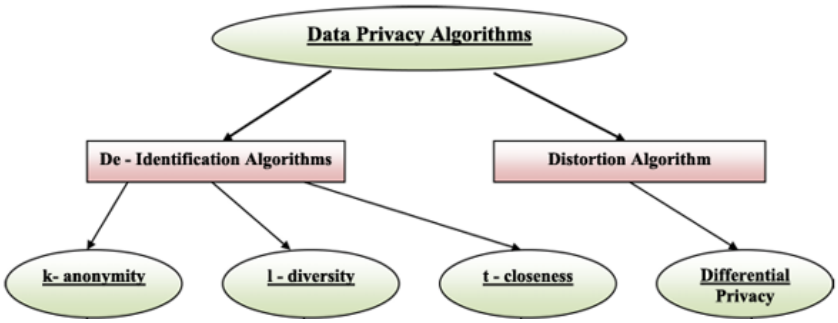
Output SDC (applied to analytical results): These techniques are applied to summary statistics and outputs such as tables or regression coefficients, and include:

- Threshold rules (e.g., minimum cell counts)
- Rounding and perturbation of cell values
- Suppression of sensitive cells in tabular outputs
- Dominance rules to prevent disproportionate influence of small subgroups

The challenge with implementing SDC lies in the trade-off between disclosure risk and data utility. Excessive masking can render datasets analytically useless, while insufficient masking can compromise individual privacy (Garfinkel, 2015). Therefore, selecting the appropriate SDC method requires a contextual risk assessment, often guided by formal risk-utility frameworks or privacy risk metrics (Templ et al., 2022).[4.12 4.13 4.14 4.15]

4.3.3. Data Privacy Methods

Data privacy techniques employed in clinical research and real-world data analysis can broadly be categorized into two major methodological classes: distortion-based methods and de-identification-based methods. Distortion methods, such as differential privacy, data perturbation, and noise addition, deliberately modify data values to obscure individual-level information while preserving statistical utility (Dwork & Roth, 2014). These approaches offer strong formal privacy guarantees, especially in large-scale data environments, but often require careful tuning to balance privacy with analytical accuracy (Abowd, 2018). In contrast, de-identification methods focus on removing or masking direct and indirect identifiers (e.g., names, birth dates, ZIP codes, and rare diagnoses) to reduce re-identification risk. Techniques such as pseudonymization, k-anonymity, l-diversity, and data generalization fall under this category (Sweeney, 2002; Machanavajjhala et al., 2007). While de-identification is widely used in regulatory compliance (e.g., under HIPAA and GDPR), it can be vulnerable to linkage attacks when auxiliary information is available (Ohm, 2010). In practice, a hybrid approach—combining statistical distortion with structural de-identification—is often necessary to meet both privacy protection and data utility requirements (El Emam & Arbuckle, 2013).[4.16 4.17 4.18]



Source: Biswas, S., Fole, A., Khare, N. et al. Enhancing correlated big data privacy using differential privacy and machine learning. J Big Data 10, 30 (2023).

- *The k-anonymity* method is based on obscuring individual identities in a dataset by grouping at least *k* similar individuals and suppressing identifying attributes. This reduces the risk of disclosing personal information about any individual within the group (Sweeney, 2002).

- *The l-diversity* method is an extension of the k-anonymity method, which enhances privacy in datasets by reducing the granularity of data representation and providing protection against attribute disclosure. For example, each group of *k* individuals should contain at least *l* different medical conditions to prevent inference about sensitive attributes (Machanavajjhala et al., 2007).
- *The t-closeness* model is an extension of the l-diversity framework that further improves privacy protection by considering the distribution of sensitive attribute values within each anonymized group. It ensures that the distance between the distribution of a sensitive attribute in any group and the distribution of that attribute in the overall dataset does not exceed a threshold *t*, thereby limiting the risk of attribute disclosure (Li et al., 2007).
- *The Differential Privacy (DP)* is a formal privacy framework introduced in theoretical computer science that ensures the risk of identifying an individual remains low, even with access to auxiliary information. An algorithm satisfies DP if the presence or absence of a single individual's data does not significantly alter the output of a query. It provides quantifiable privacy guarantees and supports data sharing through six key properties, including resistance to linkage attacks, composition, and post-processing immunity (Dwork et al., 2006; Dwork & Roth, 2014).

The integrity and trustworthiness of clinical and real-world evidence hinge on the seamless integration of data provenance, ethics, and data privacy. Data provenance ensures transparency and traceability of data from its origin through all stages of processing and analysis. This traceability is essential for upholding ethical standards, including informed consent, fairness, and respect for persons. In turn, data

privacy safeguards individual rights by applying rigorous protections to sensitive information, reinforcing both ethical obligations and the credibility of the data's lineage. Together, these three pillars form an interconnected framework that supports responsible data stewardship, enabling high-quality, ethically sound, and privacy-compliant research. [4.19 4.20 4.21 4.22 4.23 4.24 4.25]

5. Vendor Engagement

Vendor engagement (VE) in the context of RWD and EHR (or Electronic Medical Records) is the collaboration between various healthcare organizations (such as hospitals, clinics and medical practices) and vendors that provide EHR systems. Vendor engagement is essential for successful EHR implementation and ongoing use in healthcare organizations. This partnership ensures that EHR systems meet the needs of healthcare providers and improves patient care.
Here is a list of a few EHR vendors:
* [EpicCare](https://www.epic.com)
* [Oracle Health EHR] (https://www.oracle.com/health/clinical-suite/electronic-health-record/)
* [athenahealth](https://www.athenahealth.com)
* [eClinicalWorks](https://www.eclinicalworks.com)
* [Veradigm] (https://veradigm.com/)
* [MEDITECH] (https://ehr.meditech.com/)

5.1. Selection

Healthcare organizations engage with EHR vendors to select EHR systems that best fit their needs. This involves evaluating the products, features, pricing and support services.[5.1,5.2,5.3,5.4,5.5]

5.2. Implementation

This process involves data migration, software installation, customization to meet specific workflow requirements, and staff training. The customization could be to support specialized clinical workflows, integrate with other software solutions or implementing additional features and modules.

5.3. After Implementation

Vendor engagement continues after the EHR system is implemented. Healthcare organizations rely on vendors for ongoing technical support, software updates and maintenance to ensure that the systems function smoothly and remain compliant with regulatory requirements.
Feedback and collaboration between healthcare organizations and the vendor is crucial. Feedback on usability, functionality and performance of the EHR system can inform future updates and improvements for the vendor.

5.4. Regulatory compliance

Regulatory compliance in the context of EHRs typically includes adherence to laws and regulations aimed at protecting patient privacy, ensuring data security, and promoting the interoperability of health information. Some of the key regulations that healthcare organizations and EHR vendors need to comply with include:

- Health Insurance Portability and Accountability Act (HIPAA)
- HITECH Act
- Promoting Interoperability Program
- Other regulations by FDA

among others.
In the next section, we introduce two case studies from 2 sponsors that cover some of the topics above.

5.5. Case Studies

5.5.1. Case study 1: Vendor Engagement in a Real-World Data Study

Background

In a real-world evidence trial focused on analyzing dual combination therapies used in the treatment of hypertension within a multinational cohort, vendor engagement played a crucial role. This case study illustrates key aspects of vendor engagement in RWD and RWE studies without promoting any specific organization.

Importance of Vendor Engagement

Vendor engagement is critical for several reasons:

- Access to diverse and large-scale data: Vendors provide access to extensive healthcare data

from various regions and practice settings, which is essential for conducting large-scale, multinational research studies.

- Standardization of data: Collaboration with vendors enables the use of standardized data models, ensuring consistency and comparability across studies and databases.
- Regulatory compliance: Vendors often have established protocols for managing data in compliance with regulatory requirements, ensuring patient privacy and data security.
- Technical support and expertise: Vendors offer technical support and expertise in data management and analysis, facilitating more efficient and effective research processes.
- Application of advanced technologies: Vendors may apply AI solutions across the product lifecycle, bringing precision, speed, and scale to various stages of the research process.

Implementation

Vendor engagement typically begins at the early stages of a research project and continues throughout the study to ensure ongoing access to data, technical support, and compliance with emerging regulatory requirements.

- Initiation and planning: The research team identifies the need for external data or analytics capabilities and reaches out to potential vendors.
- Negotiation and agreement: Terms of access, use of data, costs, and compliance with privacy laws are negotiated and formalized through data use agreements or contracts.
- Data access and standardization: The vendor provides access to data, often involving data transformation to fit standard models.
- Ongoing support and collaboration: The vendor offers technical support, training, and consultation throughout the project.
- Compliance and ethical oversight: Both parties ensure that the project complies with ethical standards and regulatory requirements through established frameworks.

Monitoring and Assessment

Effective monitoring and assessment of vendor engagement ensure that the collaboration meets research objectives and adheres to agreed standards and regulations. This process involves:

- Performance metrics: Establishing clear performance metrics related to data quality, timeliness of data delivery, and adherence to the project timeline and budget.
- Regular updates and meetings: Scheduling regular meetings with the vendor to review project progress, discuss challenges, and adjust plans as necessary.
- Compliance checks: Regularly reviewing processes and data handling practices to ensure compliance with regulatory requirements and data security standards.

- Technical support and issue resolution: Monitoring the responsiveness and effectiveness of the vendor in providing technical support and resolving issues.
- Data quality and utility: Assessing the quality of the data provided by vendors and its utility in meeting research objectives.
- Satisfaction surveys: Conducting surveys or interviews with the research team to gather feedback on their experience working with the vendor.
- Cost-effectiveness: Analyzing the cost-effectiveness of the engagement, considering the value of the services provided in relation to the cost and impact on the project's budget.

5.5.2. Case study 2: Vendor engagement process for the FDA approved RWE study

In this case study, we are going to look at how vendor engagement is practiced in a typical pharmaceutical organization. More specifically, we look closely at vendor selection strategies, RWD collection and management, data standardization across different sources, and the pivotal role of programming and validation in ensuring data quality and analysis readiness.

Vendor selection and data collection

Vendor selection for data sourcing depends heavily on the nature of the study analysis. A multidisciplinary team, including members from medical affairs, real-world data science, and procurement, collaboratively assesses and selects vendors. This selection process is driven by study objectives, data availability from licensed sources, and the need for external data acquisition versus in-house capabilities.
In studies where patient data is collected for research purposes, particularly those with less stringent budget constraints, the company has encountered challenges in consistently capturing and abstracting the required data. This variability is due to the absence of a universal data model like CDISC standards. However, efforts are underway to improve data capture frequency and standardization. Currently, each study may rely on unique data models, necessitating the creation of tailored programs by programmers to standardize and analyze the collected data.

Data access and standardization

In the real-world data landscape, standardization levels differ significantly from those in clinical trials. Clinical trials adhere to stringent standards primarily for regulatory acceptance and data integrity assurance. In contrast, claims databases prioritize data standardization to facilitate billing processes for healthcare providers. Electronic Medical Records (EMRs) have their own set of standardization principles, dictating what data should be collected, though specific table structures and field formats can vary widely.

The **data dictionary** serves as a critical tool for understanding and structuring EMR data. It typically includes structured data elements such as table names, field descriptions, data types, lengths, and formats. This dictionary aids in data interpretation and ensures compliance with privacy regulations. Meanwhile, various data models like OMOP and Sentinel's common data model play crucial roles in standardizing and harmonizing data from multiple sources, streamlining the analysis process.

Developing **programming specifications** for data analysis is a nuanced process, tailored to each data source's unique characteristics. Protocol and Statistical Analysis Plans (SAPs) guide this process, outlining analytical objectives and methodologies. Flexibility is essential in real-world data analysis, as unforeseen insights often necessitate adaptive approaches to data interpretation and programming.

The **operational manual** or user guide for databases provides comprehensive insights into database utilization, including enrollment, expenditures, demographics, and clinical procedures. Each database within the system is detailed, specifying the tables and content available. For example, Medicaid databases encompass diverse information related to patient care coverage, services, and clinical outcomes, organized into specific data tables based on predefined client specifications.

Monitoring and validation

External data studies require meticulous oversight by cross-functional stakeholders to monitor progress, identify anomalies, and address data-related issues promptly. Routine meetings ensure alignment with study milestones and facilitate timely interventions to maintain data integrity.

Data validation tools such as SAS, R, and Python play a pivotal role in ensuring data accuracy and quality. They enable early detection of errors and adherence to predefined validation rules and edit checks. These tools are instrumental in preparing clean datasets for analysis, verifying data completeness, and identifying any anomalies or missing values.

Summary

In summary, the utilization of (RWD) in clinical research and healthcare analytics involves navigating a landscape of diverse data sources and varying standardization requirements, and vendor engagement is a crucial part of having high quality healthcare data. The challenges of data collection, particularly in studies with budget constraints, highlight the importance of developing tailored data models and programming specifications to standardize and analyze data effectively. Vendor selection is a strategic process driven by study objectives and the availability of licensed data sources. Data dictionaries and standardizations play a crucial role in structuring and interpreting electronic medical record (EMR) data, ensuring compliance with privacy regulations. Beyond data collection and standardization, effective monitoring and assessment of vendor engagements, exemplified by collaborations with CROs in research projects, are critical. This involves establishing clear performance metrics related to data

quality and timeliness, scheduling regular updates and compliance checks, and evaluating aspects such as data utility, timeliness of delivery, and cost-effectiveness. Such practices ensure that research collaborations meet objectives, adhere to standards and regulations, and ultimately enhance the value and reliability of real-world data for informed decision-making in healthcare.

6. Fit for Purpose Assessment

Before using real-world data (RWD) in research, sponsors must assess whether the data is appropriate for the specific research question. This "fit for purpose" assessment or “feasibility analysis” ensures that RWD is both relevant and reliable for use in a clinical investigation. Relevance refers to the availability of data for key study variables—such as exposure, covariates, and outcomes—and whether it includes a sufficient number of representative subjects. Reliability refers to the accuracy, completeness, and traceability of the data. Without ensuring these factors, conclusions drawn from the study may be flawed.

The fit for purpose assessment is a critical step for determining whether RWD is suitable for a study. Researchers need to evaluate the data sources for potential biases, gaps, and limitations. For example, while injectable or intravenous medications are often captured accurately in RWD, oral or inhaled drugs may be harder to track. This means a data source that works well for one study might not be appropriate for another, even if the studies and/or data sources appears to be similar on the surface.

After identifying the appropriate study design sponsors evaluate RWD sources to determine whether they are reliable and relevant enough to answer the research question. However, unlike randomized controlled trials (RCTs), which minimize bias through randomization, studies based on RWD require alternative strategies to control for observational bias when evaluating sources of RWD to generate valid real-world evidence (RWE) within the context of a clinical investigation.

6.1. Scenario: Designing a Study Using RWD

Suppose researchers are evaluating the efficacy of a new medication based on RWD sources. Before collecting data, they define a **Hypothetical Target Trial (HTT)** based on a double-blind, randomized controlled trial. The intent to emulate the HTT forces sponsors to account for known confounding

factors such as patient demographics, medical history, and treatment protocols because these important factors would otherwise be independently and identically distributed across treatment arms due to randomization. For example, if treatment arms have imbalanced age groups, the results could skew the perceived effectiveness of the intervention.

6.2. Defining a Hypothetical Target Trial (HTT)

Defining a HTT helps researchers anticipate potential sources of bias that could affect the study results, and help programmers identify those data sources and detect their presence. Both clinical and statistical expertise is needed to clearly articulate the key design elements required to evaluate RWD sources to determine whether they are fit for purpose:

1. **Estimator**
    - Define the statistical methods used to test the study's null hypothesis.
2. **Treatment Group "Assignment"**
    - Specify who will receive the treatment and who will serve as the control group.
3. **Time Zero**
    - Determine the reference point from which changes or outcomes will be compared.
4. **Length and Frequency of Follow-up**
    - Set the duration of participant observation and how often data will be collected during that period.
5. **Sample Size**
    - Ensure that the study has an adequate number of participants to generate reliable results.
6. **Inclusion/Exclusion Criteria**
    - Clearly define who is eligible to participate in the study and who is not, based on the

research objectives.

7. **Threats to Validity**
    - Identify any factors that could affect the accuracy of the study, and devise strategies to address them.
8. **Secondary Outcomes**
    - List additional metrics beyond the main study goal, define how they will be measured, and explain their relevance.
9. **Key Subgroups**
    - Identify subgroups within the study population that are critical for understanding treatment effects.
10. **Confounding Variables**
    - Recognize potential factors that might distort the relationship between treatment and outcome, and explain how these will be controlled.
11. **Rationale for Confounder Selection**
    - Justify why specific confounders are selected for control, with evidence supporting their inclusion.

Once the HTT is defined, sponsors can assess whether the data sources available in the real world adequately capture the necessary variables and whether those data points are reliable. If the data falls short—for example, if adherence rates are not captured accurately, or if there are discrepancies in how treatments are documented—sponsors may need to revise their study design or seek additional data.

This iterative process ensures that the data used in the study is fit for the purpose of answering the research question of the clinical investigation, with all decisions to include or exclude RWD from analysis made based on objective evidence. The documentation produced throughout the process includes data definitions and study-specific data quality checks to makes it easier for reviewers to replicate the study and assess its validity.

7. Analysis and Submission

7.1. Statistical Considerations

RWE studies and RCTs differ significantly in terms of study design, control over variables, and handling of biases and confounding. Both approaches have unique strengths and limitations, and researchers must carefully consider these factors when interpreting study findings and making evidence-based decisions in medicine.

Observational nature: RWE studies are typically observational in nature, meaning they involve the analysis of data collected from routine clinical practice, electronic health records (EHRs), administrative claims databases, or registries. Data is collected without intervention or manipulation by researchers.

Lack of randomization: Unlike RCTs, RWE studies do not involve random assignment of participants to different treatment groups. Patients receive treatments based on real-world clinical decisions made by healthcare providers.

Confounding and biases: In RWE studies, confounding is a major concern due to the lack of randomization. Patients may differ systematically based on factors that influence both treatment assignment and outcomes. Controlling for all potential confounders is challenging, and residual confounding can bias results. In addition, RWE studies are susceptible to various biases such as selection bias, information bias, and confounding bias. Bias can arise due to differences in patient characteristics, incomplete or inaccurate data, and unmeasured variables affecting outcomes.

7.2. Confounding and Biases

Confounding occurs when the relationship between an independent variable (such as a treatment or exposure) and a dependent variable (an outcome) is distorted by the presence of an additional factor (a confounder) that is associated with both the independent variable and the outcome. In simpler terms, confounding arises when a third variable influences both the exposure and the outcome, making it difficult to accurately assess the true effect of the exposure on the outcome.

Selection bias: This bias occurs when the study sample is not representative of the target population, leading to erroneous conclusions. For example, if a study on a new medication only includes participants who are younger and healthier, the findings may not apply to older or sicker individuals.

Information bias: This bias arises from errors in measurement or assessment of exposure, outcome, or confounders. For instance, recall bias occurs when participants do not accurately remember past exposures or outcomes, leading to misleading associations.

Measurement Bias: Similar to information bias, this occurs when there are errors in how variables are measured, particularly relevant in RWD where data may not be collected with research purposes in mind and thus may lack standardization.

Attrition Bias: This can happen in longitudinal RWE studies if there is loss of participants over time, and this loss is not random but related to the characteristics of the individuals or their treatment.

7.3. Methods to address confounding in RWD

Study design: Establish clear protocols for data collection and analysis before initiating the study. This helps minimize the risk of selective reporting and data-driven decisions. Whenever possible, include appropriate comparison groups to control for confounding factors. For example, use control groups or historical comparators to assess the impact of interventions or exposures.

Statistical adjustment: Use regression models (e.g., logistic regression, Cox proportional hazards models) that incorporate multiple covariates to adjust for potential confounding variables. Include relevant patient characteristics, disease severity, and other factors known to influence outcomes.

Matching: Match treated, and control subjects based on key covariates (e.g., age, sex, comorbidities) using simple matching or propensity score matching. This creates comparable groups and reduces the impact of confounding.

Stratification: Stratify the study population based on key confounders (e.g., age groups, disease severity) and conduct separate analyses within each stratum. This allows for examination of treatment effects within specific subgroups while controlling for confounding variables.

Sensitivity Analysis: Conducting sensitivity analyses to assess how sensitive results are to changes in the method of confounder adjustment or the presence of unmeasured confounding can provide insight into the robustness of the study findings.

Post-hoc Analysis: Using post-hoc techniques to adjust for variables that become apparent as confounders only during the analysis can further refine the study outcomes.

Transparent reporting: Clearly describe the methods used for confounding adjustment, including details on covariate selection, model specification, and handling of missing data. In addition, transparently report study limitations, including potential sources of bias and unmeasured confounders. Acknowledge the inherent limitations of observational data and discuss the implications for result interpretation.

In summary, by implementing these effective strategies in RWE studies, researchers can enhance the validity and reliability of findings, despite the inherent challenges of confounding and biases. However, it's important to recognize that no method can eliminate bias, and careful consideration of study design and data quality remains essential in generating meaningful real-world evidence.

7.4. Submission of RWD: Current Regulatory Landscape

The FDA will focus on three areas when evaluating RWD submitted in support of a marketing application. These are 1. Whether RWD are fit for use 2. Whether the study design can provide adequate evidence and 3. Whether the study conduct meets regulatory requirements. These are described in detail in the *Framework for FDA's Real World Evidence Program* published in 2018[7.1].

The FDA published detailed guidance on data standards required when submitting RWD in support of a marketing application in *Data Standards for Drug and Biological Product Submissions Containing Real-World Data Guidance for Industry* (Dec 2023)[7.2]. The key takeaway described in this guidance is that RWD and data from other non-interventional study designs must be submitted using the standards documented in the FDA Data Standards Catalog. To date, this means that RWD *must* be submitted using Clinical Data Interchange Standards Consortium (CDISC) standards; thus, sponsors need to convert real-world datasets into CDISC format when submitting this information in support of a marketing application. This guidance also provides advice on mapping RWD in CDISC format and considerations for the required documentation. While the agency acknowledges that its current catalog of standards does not necessarily reflect data derived from real-world sources, it has indicated that it is considering updates, including recommendations for mapping.

The landscape for submitting RWD is still evolving. FDA guidance documents indicate that the data standards and documentation required for submitting RWD are the same as for submitting RCT data. This presents a lot of challenges for the sponsors. Most of these challenges stem from the fact that RWD, by definition, is not collected under the supervision of a protocol and by research study staff. In addition, the business cases for using RWD and RCTs differ. RWD is typically not designed or collected

for research purposes. As a result, real world databases are organized and configured in a way that makes sense to health care providers, not researchers. Furthermore, these databases utilize different data standards that contain different concepts and coding systems compared to those defined within CDISC and used in regulatory review[7.3].

As far as exchange standards are concerned, the required standard for data exchange for FDA is SAS V5 transport file format. Although this has been the predominant format for submission of clinical data, it imposes several restrictions. Limited data types, alphanumeric variable names, limited length for variable names, labels and widths to name a few. With the breadth and depth of RWD data sources, SAS V5 transport format increasingly becoming a difficult choice for sponsors and thus not fit as a viable option for data exchange. While CDISC datasets are currently exchanged in SAS V5 transport format, CDISC is not inherently tied to this format. CDISC datasets can also be exchanged using XML, JSON, or other file formats. Thus, if the FDA and other regulatory authorities agree, CDISC datasets may be exchanged using another format[7.4].

7.5. How regulatory submission process for RCT translate in RWD:

Overview of Clinical Study Data Reviewer's Guide (cSDRG) – Legacy tabulation datasets:
CDISC standard cSDRG can be created which provides information about raw datasets and terminology that benefit from additional explanation beyond the Data Definitions document (define.xml). At many instances summary of SDTM conformance findings cannot be included in this reviewer's guide because the raw legacy data has not been converted to SDTM format. The standard cSDRG template can be used for filing.
- In Data Format and Import Information section, details about XPORT files and example codes for importing the XPORT files into SAS or R need to be written.

Overview of Analysis Data Reviewer's Guide (ADRG) – Legacy Analysis Datasets:
CDISC standard ADRG can be created which provides information about analysis datasets and terminology that benefit from additional explanation beyond the Data Definitions document (define.xml).
- Like cSDRG, ADRG also cannot have summary of ADaM conformance findings, as the analysis datasets are not using CDISC ADaM format. The standard ADRG template can be used for filing.
- If R markdown coding is done for datasets and reports submission then R studio related packages need to be mentioned with the version and functions used (project specific/CRAN).

- In Data Format and Import Information section, details about XPORT files and example codes for importing the XPORT files into SAS or R need to be written.

Define.xml (for tabulation and analysis datasets):
This file has retrospective description of variable derivations.
Challenges to create RWD define.xml are: If there are any R datasets, they need to be translated to SAS datasets, to convert non-standard to CDISC-like format, modifications are necessary to adjust define stylesheet in order to accommodate longer dataset and variable names of legacy data to display legibly in order to have the same "look and feel" of define.xml v2.0.

Case Study:
Have a Pre-NDA meeting with FDA:
- Get clarification of how the FDA can perform analyses, usually they are open to work on SAS & R but would need deeper information on handling missing data or invalid data in registry and procedure for variables derivation, which can be found in SAP.
- FDA has asked for patient-level data to facilitate a complete review of the analysis results. And so, data files were submitted in native format available i.e. "Standard Analytic Files" metadata and data collection forms.
- They agreed for R markdown programs in submission but requested all R packages(versions) along with functions.
- FDA requested data dictionaries.
- There was a proposal to submit v8 xpt because the datasets contain variables and labels of length greater than 8 and 40, respectively. But FDA said it is not their policy to accept these file formats for XPTs as version 8 have limitations including increased file size, no native mechanism for support of audit trails and referencing data sources. All electronic submissions for NDA should use version 5 and follow the maximum permissible number of characters (8 characters for variable names, 40 for variable labels, and 40 characters for dataset labels) based on Study Data Technical Conformance Guide (July 2020, version 4.5.1) [7.5]. After all the discussion FDA agreed to have both v8 and v5 XPTs but with a linked mapping document, which says about the dataset names, variable names, variables labels needed to be truncated that mapped the original names to the shortened names.
- eSUB package should be sent through eCTD application so that the information is all in one place.

Submission package deliverables for RWD:
1) Legacy tabulation datasets
- aCRF
- SAS v8/v5 xpt files
- Define files
- Reviewers guide
2) Legacy Analysis Datasets

- SAS v8/v5 xpt file - There was just 1 dataset with one record per transplant recipient, derived from intermediate datasets (.R files)
- Define files
- Reviewers guide

3) Legacy Programs

a) Dataset programs

- SAS to CSV (.sas file) - Convert sas datasets to csv files
- R markdown programs (.Rmd files) - reads in csv files and creates intermediate R datasets (.Rdata files) & final analysis dataset
- R markdown reports (.html & .pdf files) - results generated by programs

b) output programs

- Rmd programs that create R functions for analysis and tables/figures(PDF reports) in .Rmd files

7.6. Submission of RWD: What the future holds

Current FDA guidance[7.2] dictates that clinical data must be submitted using the standards documented in the FDA Data Standards Catalog[7.6]. This means, for the foreseeable future CDISC will remain the de-facto submission standard despite the fact that it was custom built for supporting RCTs and not quite fit for the diverse array of RWD. This requirement results in cumbersome data transformations, non-standard variables and domains, and business and validation rules which primary apply to data collected in RCTs. The challenges of using CDISC standards for submitting RWD have been discussed in detail by Jeff Abolafia et al in their papers[7.7,7.8,7.9].

As a viable alternative Observational Medical Outcomes Partnership – Common Data Model (OMOP CDM) or HL7 Fast Healthcare Interoperability Resources (FHIR) could be considered and developed to be submission ready. There are many advantages of using either of the two. Not least, they both are made to capture RWD and other non-interventional study designs optimally. FHIR has been optimized for EHR and claims data while OMOP CDM has been tailor made for registry and other observational study designs.

However, there may need to be a paradigm shift in future regulatory submission standards wherein the argument should move away from a single submission standard to a more hybrid submission approach. This approach needs considerable effort between industry, regulators and standards authorities to work together to harmonize and develop interoperability among various data standards for future regulatory submissions.

8. Glossary

- **RWD**: **Real-world data** are data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources. Examples of RWD include data derived from electronic health records, medical claims data, data from product or disease registries, and data gathered from other sources (such as digital health technologies) that can inform on health status.
- **RWE**: **Real-world evidence** is the clinical evidence about the usage and potential benefits or risks of a medical product derived from analysis of RWD.

- **Estimator**: In statistics, an estimator is a rule or formula used to estimate an unknown quantity or parameter based on observed data. It's a method to make an educated guess about a population parameter, such as the mean or proportion, using sample data.
- **HTT**: Hypothetical target trial.
- **Length and Frequency of Follow-up**: In research studies, especially longitudinal or observational studies, the length of follow-up refers to the duration over which participants are observed or tracked after the initial assessment or intervention. The frequency of follow-up indicates how often data collection occurs during that period. Both length and frequency of follow-up are crucial for understanding the trajectory of outcomes or changes over time.
- **Sample Size**: Sample size refers to the number of individuals or units included in a study or experiment. It's a fundamental aspect of study design, as it affects the reliability and generalizability of the results. A larger sample size generally provides more precise estimates and enhances the statistical power of the study to detect meaningful effects or differences between groups.
- **SPIFD**: Structured Process to Identify Fit-For-Purpose Data.
- **SPIFD2**: Structured Process to Identify Fit-For-Purpose Study Design and Data. Framework to Generate Valid and Transparent Real-World Evidence.
- **SPACE**: Structured Preapproval and Postapproval Comparative Study Design.
- **STaRT-RWE**: Structured Template and Reporting Tool for Real-World Evidence.
- **Time Zero**: Time zero, also known as baseline, is the starting point or initial measurement in a study or experiment. It's the moment when observations or measurements begin, often used as a reference point for comparing changes or outcomes over time.
- **Treatment Group "Assignment"**: In experimental studies, particularly in clinical trials, treatment group assignment refers to the process of allocating participants to different groups receiving different treatments or interventions. It's how researchers decide who receives the treatment being tested and who serves as the control group.

9. Disclaimer:

The opinions expressed in this document are those of the authors and should not be construed to represent the opinions of PHUSE members; respective companies/organizations or Regulator's views or policies. The content in this document should not be interpreted as a data standard and/or information required by Regulatory Authorities.

10. Appendices:

Appendix 1.1: Checklist

a) Details of Clinical Study

i) Type of Clinical study Per FDA framework

☐ Interventional Study ☐Observational Studies

ii) Clinical Study Purpose[2]

☐ Randomized Clinical Trial, that uses RWD to capture clinical outcomes related to safety or effectiveness

☐ Single arm trial, that uses RWD as an external control arm

☐ Observational studies, such as observational cohort or case-control, that generates RWE intended to help support an efficacy supplement

☐ Clinical trial or observational study that uses RWD/RWE to fulfill a post marketing requirement (PMR)

b) **Real -World Data Sources and Categorization**[10.1]

☐ Electronic Health/ Medical Records

- ☐ Input data in HL7- Fast Healthcare Interoperability Resources (FHIR) platform
- ☐ Input data not in HL7 but follows proprietary standards
- ☐ Input data does not follow any specific platform or data standards
- ☐ Input data is SDTM like

☐ Not applicable

☐ Medical Claims Data

- ☐ Proprietary efforts already taken to transform and curate the data to CDISC
- ☐ No efforts taken to transform and curate the data to CDISC
- ☐ Input data is SDTM like
- ☐ Not applicable

☐ Product or Disease Registry

- ☐ Registry follows NIH Common data elements
- ☐ OHDSI OMOP CDM model
- ☐ Registry follows other proprietary data standards
- ☐ Registry does not follow any data standards
- ☐ Input data is SDTM like
- ☐ Not applicable

☐ Data Obtained from Digital Health Technologies

- ☐ Input data is SDTM like
- ☐ Not applicable

☐ Other Data Sources (Online health community, social media data, quality of life data collected from other platforms)

- ☐ Data sources follow proprietary data standards
- ☐ Data sources do not follow any data standards

c) Data Curation and Compliance Process

i) Based on the source of data, has your organization established a data curation process?

☐ Yes ☐ No

ii) Are you familiar with the routine of data migration plan to enable timely transfer of data from RWD sources?

☐ Yes ☐ No

iii) Do you have a process to transform unstructured source data into structured source

data?

☐ Yes ☐ No

iv) Is there a process in your organization to harmonize the structured data across the system?

☐ Yes ☐ No

v) Coding system mapping- Does the curation process involve dictionary term mapping? (e.g.- Systematized Nomenclature of Medicines- Clinical Terms (SNOMED-CT) to ICD-10-CM)

☐ Yes ☐ No

d) Communication with External Stakeholders

i) Based on the inclusion of RWD in the clinical study, is your business function involved in communication with regulatory bodies?

☐ Yes ☐ No

ii) For an oncology study, have you referred to the FDA QCARD initiative for communication with US FDA? [10.2]

☐ Yes ☐ No

iii) Have you decided the timepoints in clinical trial progression, when you need to have communication with regulatory bodies?[10.3]

☐ Yes ☐ No

iv) Are you familiar with the contractual terms pertaining to data format and transfer between your organization and the external vendor providing RWD?

☐ Yes ☐ No

v) Are you aware of the platform and technology that the external RWD vendor uses to transfer

RWD to your company system?

☐ Yes ☐ No

Appendix 1.2: Case Examples

Check-a-1: Type of Study

Option Selected by User- Observational Study

Implications and Decision Factors:

1) Refer to FDA guidance[10.4.] to support marketing application, sponsor may be required to schedule a type-C meeting through existing IND product.
2) Sponsor is required to do thorough documentation and annotation of programming codes pertaining to real world data.
3) Further checks need to be evaluated by understanding the data submission requirements of health authorities.

Check-b-1: RWD sources and Categorization

Option Selected by User-

Electronic Health Records: Input data not in HL7 but follows proprietary standards

Implications and Decision Factors: Involvement of and awareness about following data operations processes by statistical reporting group-

1) Data standards followed by RWD vendor
2) Traceability and Provenance assessment plan of source data.
3) Data migration plan, and use of platform (either proprietary or vendor specific)
4) Data curation plan
5) Planning for communication with health authorities regarding the nature, and source of data, and the data curation plan.
6) Inclusion of above factors in planning for data submission timelines.

Check- c: Data Curation and Compliance Process

iii) Do you have a process to transform unstructured source data into structured source data?

Option Selected by User: Yes

Implications:

1) Statistical reporting group needs to assess applicability of existing data transformation processes for selected data sources for clinical study.
2) Conduct gap assessment to ensure that existing data transformation processes can handle the source data with its existing standard and platform.
3) Evaluate and plan for any potential risks in these processes and factor in that time in planning for submission timelines.

11. References

1.1: FDA Guidance: Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products Guidance for Industry (2024).

1.2: https://pharmatimes.com/web_exclusives/Six_2020_predictions_for_real-world_evidence_1329688/

1.3: https://www.ema.europa.eu/en/documents/report/real-world-evidence-framework-support-eu-regulatory-decision-making-report-experience-gained-regulator-led-studies-september-2021-february-2023_en.pdf

1.4: https://www.fda.gov/drugs/our-perspective/fda-approval-demonstrates-role-real-world-evidence-regulatory-decision-making-drug-effectiveness

1.5: https://www.ema.europa.eu/en/guideline-registry-based-studies-scientific-guideline#current-effective-version-section

1.6: Timothy A. Yap et al. (2022). Application of Real World Data to External Control Groups in Oncology Clinical Trial Drug Development. Frontier in Oncology. www.frontiersin.org.

1.7: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-design-and-conduct-externally-controlled-trials-drug-and-biological-products

2.1,2.2,2.3: ICH M14 Guideline on general principles on plan, design and analysis of pharmacoepidemiological studies that utilize real-world data for safety assessment of medicines, May 2024.

2.4: "Tools to consider when engaging health authorities", Real World Data Solutions - TransCelerate (transceleratebiopharmainc.com)

4.1 Buneman, P., Khanna, S., & Tan, W. C. (2001). Why and where: A characterization of data provenance. *International Conference on Database Theory*, 316–330. https://doi.org/10.1007/3-540-44503-X_20

4.2 El Emam, K., & Arbuckle, L. (2013). *Anonymizing Health Data: Case Studies and Methods to Get You Started*. O'Reilly Media.

4.3 Malin, B., Loukides, G., Benitez, K., & Clayton, E. W. (2013). Identifiability in biobanks: Models, measures, and mitigation strategies. *Human Genetics*, 132, 233–245. https://doi.org/10.1007/s00439-012-1235-0

4.4 Simmhan, Y. L., Plale, B., & Gannon, D. (2005). A survey of data provenance in e-science. *ACM SIGMOD Record*, 34(3), 31–36. https://doi.org/10.1145/1084805.1084812

4.5 Tene, O., & Polonetsky, J. (2013). Big data for all: Privacy and user control in the age of analytics. *Northwestern Journal of Technology and Intellectual Property*, 11(5), 239–273.

4.6 European Union. (2016). **Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016** on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). *Official Journal of the European Union*, L119, 1–88. https://eur-lex.europa.eu/eli/reg/2016/679/oj

4.7 Government of Canada. (2000). **Personal Information Protection and Electronic Documents Act (S.C. 2000, c. 5)**. Retrieved from https://laws-lois.justice.gc.ca/eng/acts/P-8.6/

4.8 Health Canada. (2019). *Public Release of Clinical Information: Guidance Document*. Retrieved from https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance/document.html

4.9 U.S. Department of Health and Human Services. (2000). **Standards for Privacy of Individually Identifiable Health Information; Final Rule (HIPAA Privacy Rule)**. *Federal Register*, 65(250), 82462–82829. https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html

4.10 European Medicines Agency. (2022). *Clinical Trials Regulation*. Retrieved from https://www.ema.europa.eu/en/human-regulatory-overview/research-development/clinical-trials-human-medicines/clinical-trials-regulation

4.11 European Medicines Agency. (2014). **European Medicines Agency policy on publication of clinical data for medicinal products for human use (Policy 0070)**. EMA/240810/2013. https://www.ema.europa.eu/en/documents/policy/0070-policy_en.pdf

European Medicines Agency. (2010). **European Medicines Agency policy on access to documents (Policy 0043)**. EMA/110196/2006. https://www.ema.europa.eu/en/documents/other/ema-policy-

0043_en.pdf

4.12 Reference: Chu, A. M. Y., Lam, B. S. Y., Tiwari, A., & So, M. K. P. (2019). An Empirical Study of Applying Statistical Disclosure Control Methods to Public Health Research. International journal of environmental research and public health, 16(22), 4519. https://doi.org/10.3390/ijerph16224519}.

4.13 Abowd, J. M. (2018). The U.S. Census Bureau adopts differential privacy. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2867–2867. https://doi.org/10.1145/3219819.3226070

4.14 Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography Conference*, 265–284. https://doi.org/10.1007/11681878_14

4.15 Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4), 211–407. https://doi.org/10.1561/0400000042

4.16 El Emam, K., & Arbuckle, L. (2013). *Anonymizing health data: Case studies and methods to get you started*. O'Reilly Media.

4.17 Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-Diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1), 3. https://doi.org/10.1145/1217299.1217302

4.18 Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57, 1701–1777.

4.19 Sweeney, L. (2002). k-Anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 557–570. https://doi.org/10.1142/S0218488502001648

4.20 Li, N., Li, T., & Venkatasubramanian, S. (2007). t-Closeness: Privacy beyond k-anonymity and l-diversity. *IEEE 23rd International Conference on Data Engineering*, 106–115. https://doi.org/10.1109/ICDE.2007.367856

4.21 Floridi, L., & Taddeo, M. (2016). What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083), 20160360. https://doi.org/10.1098/rsta.2016.0360

4.22 Garfinkel, S. L. (2015). De-identification of personal information. *NISTIR 8053*, National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8053

4.23 Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., & de Wolf, P.-P. (2012). *Statistical Disclosure Control*. Wiley.

4.24 Templ, M., Meindl, B., Kowarik, A., & Alfons, A. (2022). *Statistical Disclosure Control in Practice: Protecting Confidentiality in Survey Microdata*. Springer.

4.25 Willenborg, L., & de Waal, T. (2001). *Elements of Statistical Disclosure Control*. Springer. https://doi.org/10.1007/978-1-4757-3454-7

6.1: ICH M14 guidelines on general principles on plan, design and analysis of pharmacoepidemiological studies that utilize real-world data for safety assessment of medicines, May 2024

6.2: RWD: Assessing electronic health records and medical claims data to support regulatory decision making for drug and biologic products – guidance for industry July 2024

6.3: A Structured Preapproval and Post approval Comparative Study Design Framework to Generate Valid and Transparent Real-World Evidence for Regulatory Decisions - PubMed (nih.gov)

6.4: The Structured Process to Identify Fit-For-Purpose Data: A Data Feasibility Assessment Framework - PubMed (nih.gov)

6.5: A Structured Process to Identify Fit-for-Purpose Study Design and Data to Generate Valid and Transparent Real-World Evidence for Regulatory Uses - Gatto - 2023 - Clinical Pharmacology & Therapeutics - Wiley Online Library

7.1: FDA Guidance (2018). "Framework for FDA's Real-World Evidence Program". December 2018. https://www.fda.gov/media/120060/download

7.2: FDA Guidance. (2023). "Data Standards for Drug and Biological Product Submissions Containing Real-World Data Guidance for Industry". December 2023. https://www.fda.gov/media/153341/download

7.3, 7.4: Abolafia, J, Ferko, S, & Holt, I (2024) "Future Clinical Data Submission Standards: CDISC, FHIR, OMOP, or Hybrid Model" Paper presented at the PHUSE US Annual Conference 2024, Bethesda, MD

7.5: FDA Guidance (2020). "Study Data Technical Conformance Guide". November 2020. https://www.fda.gov/media/143550/download

7.6: FDA Guidance. (2023). "Data Standards Catalog". April 2023. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/data-standards-catalog

7.7: 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

7.8: 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

7.9: Abolafia, J, Ferko, S, & Holt, I. (2023). "Submission Standards for Real World Data: Gaps, Limitations and Recommendations". Paper presented at the PHUSE Annual Conference 2023, Birmingham, United Kingdom. https://phuse.s3.eu-central-1.amazonaws.com/Archive/2023/Connect/EU/Birmingham/PAP_RE03.pdf

10.1: Data Standards for Drug and Biological Product Submissions Containing Real-World Data -Guidance for Industry, Draft guidance by U.S. FDA. Released on October, 2021

10.2: Oncology Quality, Characterization and Assessment of Real-World Data (QCARD) Initiative | FDA

10.3: "Tools to consider when engaging health authorities", Real World Data Solutions - TransCelerate (transceleratebiopharmainc.com)

10.4: FDA Guidance: Considerations for the Use of Real-World Data and Real-World Evidence to Support Regulatory Decision-Making for Drug and Biological Products, Final Guidance, August 2023

12. Project Contact Information:

- Dhruba Sikdar; Chi Zhang; Matt Baldwin; Elena Valkanova; Xingshu Zhu; Sherry Zhong; Yen Phan; Jingying Zhou; James Joseph; Priyadarshini Tunga; Srinath Yarasi; Sowmya Gabbula
- Use Email: workinggroups@phuse.eu

13. Acknowledgments:

The authors would like to express their gratitude to RWE Working Group in PHUSE, without their support and guidance this project would not have been possible.

The authors would also like to express gratitude to the Integrated Data Analysis and Reporting (IDAR) function in Johnson and Johnson Innovative Medicine (JJIM). The structure and content of this paper are generously borrowed from the RWE playbook that IDAR-JJIM has developed for internal use. For that the authors are highly indebted to the Johnson and Johnson team who developed the playbook.

Sincere gratitude to the *Submitting Real World Data* project led by Parag Shiralkar for providing the submission checklist.

Lastly, the authors are grateful to their respective organizations for allowing them the opportunity to work on this project.