Paper AS02

Planning and Execution of Real-World Data Evidence Analysis Based on Janssen Initial Experience

Ramesh Sundaram, IQVIA, Ontario, Canada
Nicole Thorne, Janssen Pharmaceutical Group of Companies of Johnson & Johnson, Springhouse, PA & Raritan, NJ

ABSTRACT

The usage of Real World Data Evidence (RWDE) can optimize the cost and time of drug development process, so in recent days the role of RWDE analysis are significantly increased in each stage of drug development and its decision making process, like identifying target patient population, finding right variants, gene mutation, even comparators for an investigational product, etc.

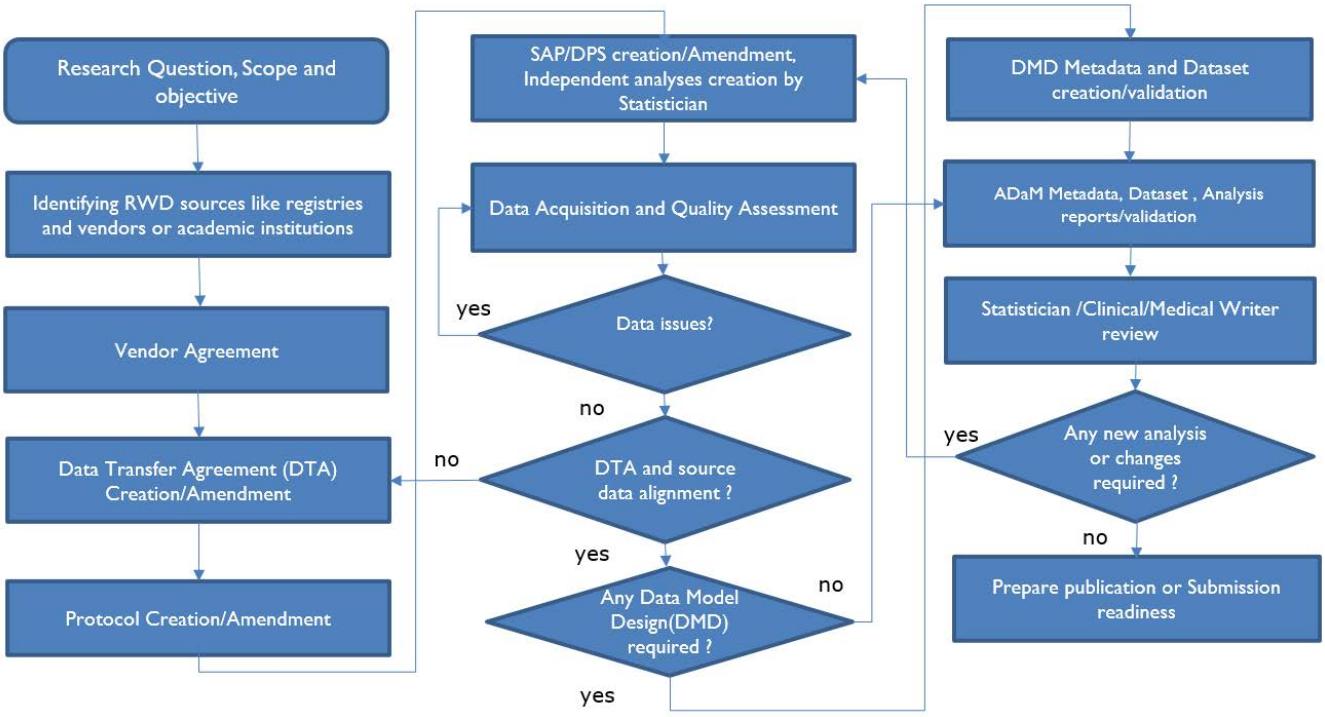
Unlike clinical studies, working with RWDE data has its own challenges, this paper demonstrates planning and execution of RWDE, like general workflow, data analysis, data Integration (RWDE & clinical data) related challenges and workarounds, overview of data review & cleaning process and a list of factors needs to consider for overall timeline estimation.

INTRODUCTION

Real World Evidence (RWE) data source can be from electronic health records (EHR), medical billings, independent research centers. Unlike Clinical trials, data from real-world sources are not robust as there are no established standards for data collection. Therefore, challenges like inconsistent and redundant data points and partial dates are quite common. Proper planning can overcome these challenges. This paper is designed from the statistical analysis, programming, and reporting perspective based on Janssen initial experience.

PROCESS FLOW

The overall RWE process flow is slightly different from the typical clinical trial flow, the below diagram describes a sample RWE process flow.



VENDOR SELECTION

Finding an appropriate data source like registries and vendors or academic institutions is the initial process, as part of this exercise at least following activities are performed, feasibility review checks like research and indication background, population size, patient entry/selection criteria etc.

VENDOR AGREEMENT

A detailed collaborative vendor agreement could be created with following sections, research title/name, requestor and/or lead researcher narrative summary, research proposal & background, objectives, patient population size, research methods, timeline, patient selection criteria, data source, high level overview of data, budget, data privacy, etc.

DATA TRANSFER AGREEMENT(DTA)

Data manager creates data transfer agreement in collaboration with cross functional team. This is the primary document for SAP creation and it mainly covers following sections, point of contacts (at least one person from each functional area), frequency of data transfer (weekly/biweekly/monthly), source file type (.csv,.txt,.xls,.xlsx), method and location of data transfer, source dataset level metadata, variable and value level metadata. This document can be amended in future i.e. in case of any metadata deviation in actual database.

PROTOCOL, STATISTICAL ANALYSIS PLAN (SAP), DATA PRESENTATION SPECIFICATION (DPS) CREATION

Vendor agreement and DTA are key inputs for protocol, SAP and DPS creation. As per health authority requirement cross functional team closely work together to finalize and sign these documents before the initial data transfer. These documents are created without seeing actual source data but based on research objective and vendor provided metadata, so to minimize amendments generic analysis sections with alternate analysis procedures are described. If requires these documents can be amended in future.

REGULATORY AGENCY TOUCH POINT

Recommended to discuss the RWE usage plan with the regulatory agency like FDA if any such submissions are required. Refer to the FDA guideline for additional details. FDA can request an additional technical meeting so ensure following things before that.

- Data Privacy laws are met, in each country
- Data traceability with focus on risk management in case (multiple) vendors supply the vendor being assessed/ qualified.
- Being specific on Data Use, Data Quality, Data Transfer Agreements
- Evaluating need to link supplier risk assessment to protocol/study risk assessment process.
- Ensuring Regulatory requirements are met for e-submission & traceability for any submitted data
- Asking for additional proof/audit if systems documentation not available for part 11 compliance

DATA AND DTA ALIGNMENT CHECKS

After initial data transfer, data manager processes (convert any non-SAS file into. sas7bdat file) and loads them into analysis area then cross functional team performs edit checks, DTA alignment, integrity checks based on objectives.

All key primary variables are thoroughly reviewed from data integrity and traceability perspective, example if progression free survival is one of the objectives, then the availability of progression date, treatment/diagnosis date, the last known alive date are required.

To ease the review process, statistician/programming team can design some SAS/R codes to perform following data checks, missing/partial dates, inconsistent data points, missing values, etc. All potential data related issues are communicated to vendor on timely basis. Regular vendor meetings are conducted to speed up the data resolution process.

In some instances, statistician creates independent analysis and reports, which can be used for clinical protocol developments, to understand indication background, to validate/review formal outputs, etc.

ANALYSIS PREPARTION AND VALIDATION

Mostly source datafiles are nonstandard, so if standard formats are required for the source data then SDTM or other Data Model Design (FHIR/Sentinel CDM/PCORNET CDM/OMOP CDM) specification/datasets and dependent ADaM specification/datasets can be created in case of any submission requirements. Otherwise all required and dependent nonstandard source variables are directly mapped to ADaM standards during this process data integrity and traceability principles are carefully reviewed.

To ease this direct ADaM creation process, datasets can be categorized into two levels tier 1 and 2. Subject level information and all required direct source variables are mapped to tier 1 datasets example ADSL, ADCM/ADLOT/ADRS. Tier 2 datasets are created based on Tier 1 variables, example ADTTE, ADEFF. ADaM metadata are designed in line with respective clinical study metadata (i.e. variable name, parameter name, label, structure, derivation rules, imputation rules), this can ease future data pooling process.

All formal reports are programmed from standard ADaM datasets and formally validated. Just like usual clinical study analysis one or two dry runs are highly recommended before final data transfer. Statistician team reviews ADaM metadata, datasets, and reports. Especially key objectives related derivations and reports are prioritized. After incorporating all statistician and cross functional findings, all final reports are transferred to common location for publication or submission purpose. For submission purpose e-sub packages are created. During analysis preparation documentation of derivation, imputation rules, leaky pipe information, etc. are vital and this can be used for ADRG preparation.

KEY FACTORS IMPACTING TIMELINE

Timeline estimation is the core part of planning process and its essential to consider associated potential factors, which could impact this. Following factors, number of data vendors, source data file type, data transfer approach/method, data issues/resolution and volume of required analysis can impact the overall RWE timeline estimation.

NUMBER OF VENDORS

Dealing with multiple vendors may need more time because each vendor can have their own data structures and formats, so optimizing/integrating them into required standards data model can require some extra time. For an example, a large sample size like 5K patients may not be possible from a single data vendor, so multi-vendor data are required.

SOURCE DATA FILE TYPE

Mostly RWE source data files are non-SAS (.csv, .txt,.xls,.xlsx,.html) and nonnormalized file formats, so conversion and validation process for each data transfer need some extra time. In the below example, the column metastatic lesion in lung had multiple recurrence timepoint values in a cell.

Patient ID	Race	Ethnicity	Gender	Birth Year	Metastatic Lesion in Lung
57899889	White	Not Hispanic or Latino	M	1945	Recurrence1:N;Recurrence2:N;Recurrence3:N;Recurrence4:N;Recurrence5:N
7280726	White	Not Hispanic or Latino	M	1927	Recurrence1:N;Recurrence2:N;Recurrence3:Y;Recurrence4:N;Recurrence5:N;Recurrence6:N;Recurrence7:Y;Recurrence8:N;Recurrence9:N

Note: SAS V5 transport format files, can save some conversion and validation time.

DATA TRANSFER

Data transfers can be in multiple batches, delay in transfer or any structural changes in source data during batch wise transfer, i.e., addition of new variables or deletion of existing ones, can delay the overall time, because this can impact the existing analysis and reporting programs, so reruns can consume extra time. Example a vendor transferred the clinical database on time but didn't transfer the Biomarker data, which had the key comparison variables.

DATA ISSUES & RESOLUTION

Amount of data inconsistency issues is directly proportional to the volume of source data. So, while working on a large RWE sample size, keep sufficient time for cleaning and resolution activity. When issues cannot be resolved, we need to establish data handling rules.

VOLUME OF REQUIRED ANLYSES

More number of analysis and reporting need more time

BIASES

Retrospective data can often subject to sampling biases that are not present in prospective studies, where participants are actively followed starting at a particular point in time. Two common situations in which sampling bias can arise, both of which are particularly relevant in external control arms analyses. The first arises when follow-up for participants in a retrospective analysis begins before those participants have satisfied all inclusion criteria for the analysis. The second arises when participants meet the inclusion criteria for an analysis at multiple different points in time. Note that there are many other similar types of bias to be aware of, including ascertainment bias, where data is collected so that some members of the target population (for example, those with more severe outcomes) are more likely to be included than other members.

TIMELINE ESTIMATION

Below table describes an approximate time estimate for each task with task owner and reviewers, for a single data vendor, where data transfer, data quality issues/resolution times are nominal, and number of required analysis datasets are 4 and reports are 35-40.

Task	Timeline	Owner	Reviewer(s)	Assumptions
Vendor Agreement	Approximately 5 to 6 weeks before the initial data transfer	Translational Research/Clinical	Clinical leadership at the Therapeutic area level, Study responsible physician, Clinical leader	Clinical leadership at the TA level must approve as part of the governance process for concept and funding
DTA Creation	Usually, a week before initial data transfer. (a week)	Data Manager (DM)	Clinical/Statistician /Programming leads	If no complex requirement like where to upload data etc..
Protocol/SAP/DPS	Usually, singed and finalized before the first data transfer	Clinical/ Statistician /Programming lead	Clinical/Medical Writer/Translational research/ Programming leads	In parallel with above task
DTA & source data alignment, Data issues checks	Within 2 weeks from first transfer	Programming lead	Clinical/Statistician leads	This includes data integrity, traceability, list of source datasets, variables, parameters etc.
Statistician independent Analysis datasets & reports	Within 4-5 weeks from SAP	Statistician lead	Clinical/Programming leads	
Data Model Design (DMD) Metadata creation (if required)	3-4 weeks from first transfer	Programming lead	Independent programming validation/ Statistician review	If SDTM required, only required domains are created (10 to 13 core domains, 8 to 10 supplement domains), 100% CDSIC compliance is not feasible 3-4 FT programming resource
Data Model Design (DMD) Dataset creation/validation (if required)	4-5 weeks from Metadata	Programming lead	Risk based Independent programming checks/ Statistician review	
Data Model Design (DMD) e-sub preparation (if required)	2 weeks	Programming lead	Independent programming review/ Statistician review	eCTD package preparation
ADaM Metadata/Programming/Validation	within 2-3 weeks from DPS	Programming lead	Independent programming review/ Statistician review	4 ADaM datasets, 3 FT programming resource, (ADaMs can be started once the first draft of SDTMs are ready, when SDTMs are handled by the independent team)
TLGs Programming/Validation	2-3 weeks from ADaM	Programming lead	Independent programming review/ Statistician /Clinical/Medical Writer review	40 reports (6 to 7 unique layouts) 3 FT programming resource (TLGs can be started once the first draft of ADaMs are ready)
ADaM e-sub preparation (if required)	2 weeks from TLGs	Programming lead	Independent programming review/ Statistician review	eCTD package preparation (p21 checks, Define.xml, ADRG, Supplemental data definition)

In a summary, for 4 ADaM datasets and 35-40 analysis reports without standard source datasets (SDTM) it requires approximately 100 days for 3 to 4 FT programming resource and approximately 136 days with standard source datasets (SDTM) datasets.

Multiple reruns are required when source data are transferred in batches,

- SDTM rerun/revalidation requires 5-6 days, provided no structural changes in source data.
- ADaM/TLGs rerun/revalidation requires 3-4 days, provided no structural changes in source data.

CONCLUSION- CHALLENGES AND WORKAROUNDS

Retrospective RWE data collection model don't have much control on patient enrollment, data cleaning, processing, and transformation process, so following challenges are very common. But involvement of Stats and Programming team at the initial planning stage of RWE can overcome these Challenges.

Below table describes, common challenges and respective possible workarounds.

Challenges	Workarounds
Non normalized database • Redundant information • Inconsistent information • Not following BDS structure for vertical data, etc. Nonstandard database (Non CDISC)	Ask for data in a standard data model: • FHIR V4.0.1 Sentinel CDM V6.0.2 • PCORNET CDM V4 • OMOP CDM 5.2 • SDTM V3.3 Note: Probability of getting SDTM format is less likely
Source file type • Non-SAS files • SAS incompatible variable names, special characters like nonprintable characters	At vendor agreement level, demand SAS V5 transport compliance source files
Data integrity/traceability issues • Variables without its dependent source • Example a data vendor transferred survival data with direct OS durations without respective death date, last known date, diagnosis date, treatment start date	Detailed review of Data Transfer Agreement is highly required at the initial stage. i.e., verify and make sure all dependent and required variables are properly defined at Vendor agreement/DTA level.
Imbalance comparison groups • Number of expected subjects can be very less for a few groups, so drawing clinically/statistically meaningful results can be non-feasible	Frequency distribution of a few key variables are recommended before the vendor agreement approval/signoff
Pseudo-anonymous date shift • To preserve data privacy and patient identity, a few vendors are encrypting the source date variables by using pseudo-anonymous shift and these shifts can lead to future dates, which can raise data quality related questions example date of death =10JAN2030	On or before vendor agreement approval, review data privacy terms and date handling methods, get them fixed if anything inappropriate for the analysis scope
Clinical Study data and RWE data pooling	Pooling source database is not easy and appropriate approach ADAM pooling is more convenient approach. Note: Proactively plan & prepare all RWE ADaM structure, Metadata/attributes consistent with respective clinical study.
Partial or batch wise data transfer • A few data vendors can transfer data in batches • Sometime noncumulative datasets, this can lead to extra programming and validation task	Retrospective data collection process can be ongoing, so we can't avoid batch wise data transfer, but it's good to accept cumulative/consolidated database for every transfers. Note: If needed dummy data can be created to overcome data transfer delays
Data issues/resolution • Inconsistent data points • Redundant data points • Partial dates, etc.	Data issues are very common • Design a few simple data checks/edit checks programs for all analysis related source variables. • Communicate vendor(s) on all finding/data issues on a timely manner • Have regular meeting with data vendor(s) • Plan one or two dry runs before the final data transfer., this can avoid last minute hiccups • When data issues cannot be resolved, add some data handling rules

Challenges	Workarounds
Protocol/SAP/DPS amendments • Data handling and analysis sections are created using Vendor agreement and DTA, i.e. before seeing original data.	Generalized analysis sections can minimize number of amendments and Define alternate analysis methods/approach • For an example, for survival endpoints define both usual survival and delayed entry model, i.e. in case of Survival data with left truncation bias issue

ACKNOWLEDGMENTS

We would like to acknowledge the following cross-functional colleagues who provided support and guidance for the work referenced in this paper.

Sangeeta Bhattacharya
Asha Mahesh
Simon Walsh

RECOMMENDED READING

https://www.fda.gov/media/124748/download
https://catalogofbias.org/.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Ramesh Sundaram
IQVIA
535 Legget Dr
Kanata, ON K2K 3B8
Canada
ramesh.sundaram@quintiles.com

Nicole Thorne
Janssen
1400 McKean Road
Springhouse, PA 19477
Work Phone: 215-793-7430
nthorne@its.jnj.com