Paper RE09

From Roadblocks to Roadmaps: Overcoming Challenges in Registry Data Reporting

Neha Srivastava, Fortrea Inc., Durham, United States of America
Sivasankar Konda, Fortrea Inc., Durham, United States of America

ABSTRACT

Year after year, Real World Evidence (RWE) remains a hot topic in the world of data analysis. Despite advancement of guidance documents on RWE from regulatory bodies, the RWE programming community relies on the experience from precedents and pilots for best practices. Registries are emerging as a new gold standard for real world research; however, real-world data presents significant challenges in reporting and analysis. Registry programmers often work in non-standardized settings with diverse datasets. Global registries bring humongous unexpected missing data making data completeness checks essential before proceeding with any analysis. Rare disease registries bring large amount of genetic testing data which is still a niche area for majority of the programmers. Facing these challenges with no survival guide in hand feels like driving a 1976 Impala on 2025 roads. The paper addresses practical difficulties of working with registry data and propose potential solutions to improve programming practices amidst these challenges.

INTRODUCTION

The clinical registry is an organized system that uses observational study methods to collect clinical and other data to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure, and serves one or more predetermined scientific, clinical, or policy purposes. Properly designed and executed registries can provide a real-world view of clinical practice, patient outcomes, safety, and comparative effectiveness.
Registry reporting is often not specifically designed for regulatory submissions. It comprises of various statistical analyses to support the manuscripts, publications and presentations depicting the purpose and mission of registries. The absence of clear reporting guidelines, unlike for the clinical trial studies, can tempt programmers to dive straight into the work without adopting a well-thought-out approach. But strategic planning and thorough preparation before engaging with real-world data can be a transformative experience for the registry programmers.

"Give me six hours to chop down a tree and I will spend the first four sharpening the axe."
– Abraham Lincoln

The essence of preparation cannot be highlighted better than this quote. In the context of registry data, this paper offers essential tips and best practices to maximize the effectiveness of those initial hours, which will lead to a smoother and more streamlined registry reporting process.

OPTIMIZING REPORTING WITH A STRUCTURED FOLDER APPROACH

Reporting might seem relatively simple when starting with the first analysis request of the registry. When we began working on the first ever analysis, with no reference registry study available, the programmers focused on creating Analysis Datasets (ADs) and generating the necessary tables, listings, and figures (TLFs) using the straightforward traditional workflow: RAW Data → Analysis Datasets →TLFs

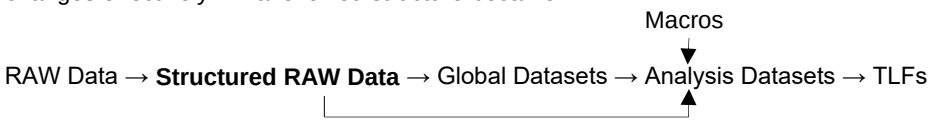
After working on couple of analyses, we realized many variables had already been derived in previous analyses. This led to a more efficient strategy: creating Global Datasets that would run each time new data was available, allowing common variables to be reused across multiple analyses. The new structure became:
RAW Data → **Global Datasets** → Analysis Datasets → TLFs

As more analyses kept coming, we found that derivations from one analysis were required in others, and these variables were not part of Global Datasets. To address this, we began using macros for frequently used and lengthy derivations. These macros were then freely utilized across multiple analyses, ensuring efficiency and consistency. The updated workflow was:

Macros
RAW Data → Global Datasets → Analysis Datasets → TLFs

With this strategy in place, we felt confident that our programming practices were robust, and our folder structure worked well for any new analysis.
Despite the progress, new challenges inevitably arose. As registries are long term and can span from 3 to 15 or more

years, changes such as protocol amendment and CRF updates are expected. A new version of the CRF was introduced a few years after the registry initiation. With the new CRF, significant updates were required in the Global Datasets and Analyses Datasets for all analyses. Now this called for an entirely new approach to handle the new looking data. We discovered that adding an additional layer of Structured RAW Data would help manage these changes effectively. Final evolved structure became –



This reporting folder set-up ensures consistency with same derivations used across different analyses and can handle protocol amendments resulting in CRF updates. By following this organized approach, we were able to streamline our work and adapt to evolving challenges.

RECOMMENDATION 1: Think ahead, brainstorm the potential changes at the initiation of the study, and set up your folder structures with these expected changes in mind. Always prioritize reusing derivations through the use of macros. While it may be tempting to seek shortcuts, especially in the absence of established guidelines, these shortcuts can ultimately lead to significant setbacks, wasting both time and effort in the long run.

STANDARDS ARE THE WHEEL, DON'T REINVENT!

Handling real world raw data can be quite challenging, but by creating Structured Raw and Global Datasets, we can organize the data in a way that is more readable and easier to manage. We recommend using Study Data Tabulation Model Implementation Guide for Human Clinical Trials (SDTMIG) as a reference to design the Structured Raw datasets, and the Analysis Data Model Implementation Guide (ADaMIG) to design the Global Datasets and Analysis Datasets. The variable names, labels, and lengths specified in the Implementation Guides can be utilized while designing the datasets.

Registries typically contain disease-specific data, and many rare disease registries also include extensive genetic testing data. We designed our genetic testing data in accordance with SDTMIG-PGx v1.0 couple of years back, although this has been deprecated and integrated into SDTMIG v3.4. In December 2023, Clinical Data Interchange Standards Consortium (CDISC) published the Therapeutic Area User Guide (TAUG) for rare diseases, providing guidance on designing SDTM and ADaMs. It is highly recommended to refer to TAUGs wherever applicable. If a TAUG for specific disease of interest is not available, consider whether any symptoms or characteristics data related to the disease can be mapped as per any existing TAUG. For example, we utilized the Kidney Transplant TAUG for one of the rare disease registries.

For Analysis Datasets and RAW data that do not have a direct reference in the available CDISC guidelines, create specifications that align with variable names outlined in the Implementation Guides. For instance, standard result variables should end with '--STRESN,' and day variables should be suffixed with '--DY.' This approach is like creating custom domains according to the SDTMIG, ensuring that all data is organized, readable, and standardized at its best.

RECOMMENDATION 2 – Scan all available CDISC guidance, including the SDTMIG, ADaMIG, and any relevant TAUGs. If there is no exact match, find the closest applicable TAUG. As Real-World Data Programmers, staying up to date with CDISC standards is crucial for effectively and confidently managing the data. Leverage these standards wherever possible, from data organization to standard variable names.

THE SAGA OF MISSING DATA

At this stage, now that we have set up the folder logistics and have a plan to design the datasets, lets investigate what happens when the data arrives.

First impression: we have hundreds of raw datasets. Whoa! The notion that ***"Data is the New Oil"*** has been around for quite some time, generally credited to mathematician Clive Humby. However, when you look at this data, it feels like an untapped reservoir.

Since registries are observational in nature, they do not dictate patient visits or mandatory lab tests. This may result in missing or incomplete data due to various factors like missed visits or inconsistent documentation. These factors, among other potential issues, result in huge missing data in the registry database. Understanding the extent of missing data and the likely impact on the ability of the registry to meet its objectives is critical before working on any analyses.

CRITICAL CORE VARIABLES

It is time to tap the reservoir and identify the Critical Core Variables. Critical Core Variables are the minimum set of data expected for a patient to be included in the registry and eligible for analysis. It is essentially study population criteria, as well as additional information on disease-specific biomarkers, signs, symptoms and severity. These include variables important for analyses of safety and effectiveness and are of specific interest to the health authorities. Critical Core Variables are prioritized for data entry, cleaning and reporting.

Like a convention, do not skip the protocol and SAP review meeting with programmers. Additionally, review the aCRF to understand what the raw data would look like. This will help the team identify the Critical Core Variables required for reporting. Critical Core Variables should be measurable and actionable. Make a list of these variables and have it reviewed and approved by study epidemiologist(s) or lead(s). Collating these Critical Core variables familiarizes the programming team with the data and registry endpoints. These Critical Core Variables then serve as input for the Data Completeness and Data Availability report.

DATA COMPLETENESS REPORT

Data Completeness report is a tool to evaluate the completeness of data from Registry CRFs. It is created to identify omissions in the database and communicate them to the sites. The report shows CRF completeness of Critical Core Variables by providing metrics (percentages) of data completeness globally, by country, and by site. The report considers all data fields marked as Unknown, Not Done and Not Available and uses CRF-based time periods which allows to communicate with sites efficiently. It provides answer to the question - Are the sites entering all patient results available, by providing statistics to 3 below data scenarios –

- The patient has result available at time point of interest
- The patient does not have result, and site has indicated that response is “No” “Unknown” or “Not Done”
- The patient does not have result, and site has not indicated that result is “No” “Unknown” or “Not Done”

The report helps Identify sites with patients who have missing data when the site has not indicated whether or not the data is truly available for the time period of interest. Alternately, the data review may identify issues with a particular site, such as high rates of missing data or loss to follow-up for certain sites.

The report was also be evolved by adding consistency checks. e.g., a “yes” answer to a question at one point and “no” to the same question at another. This can include questions related to disease symptom and characteristics. Checks for out-of-range data were also added (e.g., a 150-year-old patient and a pediatric patient with future date of birth).

RECOMMENDATION 3 – Many global registries face issues due to humongous missing data. To address this, create a data completeness report to identify what’s missing and what’s incorrect. Collaborate with the data management team to discuss these issues and explore possible causes. Setting up monthly meetings with data management team has proven very effective in handling data issues and ensuring faster resolution, thereby enhancing data quality. Based on the findings of Data Completeness report, yearly goals can also be set to improve data completeness of Core Critical Variables.

DATA AVAILABILITY REPORT

Data Completeness Report gives insight on the completeness of CRF data but cannot assess how much data is actually available for analysis as it includes ‘Not Done’ and ‘Unknown’ results.

Data Availability Report, on the other hand, evaluates amount of data that can be used in analysis. It is based on standard analysis datasets, accounts only for analyzable results, and uses defined analysis periods. The main objective of the Data Availability Report is to evaluate the availability of data from Registry CRFs that can be used in the analyses. It shows data available for analysis, aggregated by treatment, age groups or other subgroups of interest. This report excludes data marked as ‘Not Done’, ‘Unknown’, ‘Not Available’, as well as outlier results, and it uses date-based period definitions and analysis time windows. The report includes metrics showing data for patients with baseline data and assessments available at the timepoints of interest.

Additionally, the Data Availability report helps reduce change requests from reviewers. After working on the first few analyses, we observed that reviewers often returned with change requests to use specific sets of data under certain conditions, leading to considerable rework and delays from programming team. However, the Data Availability report provided the reviewers with an early view of the data, allowing them to craft more meaningful analyses and thereby reduce the iterations of change requests.

RECOMMENDATION 4 – Registry data reporting often involves several exploratory analyses, which can lead to numerous change requests. However, Data Availability Reports have proven to be highly effective in providing reviewers with a preview of the data, allowing them to decide which numbers should be used or whether it's the right time to proceed with the analysis, and hence avoiding rework and delays from the programming team. Data Availability reports are essential and become easier to generate once Data Completeness reports are set up.

THE FINAL PIECE: UNDERSTANDING YOUR CONTRIBUTION

Many registry programmers often feel lost as they work on one analysis after another without a clear end goal in sight. This usually happens because they aren't working toward specific milestones like database lock, or one particular submission. Instead, they focus on different types of analyses to support various manuscripts, conference presentations, posters, and publications, without understanding the final impact of their work. It’s crucial for study leads to communicate the overall goals to the programmers and, ideally, share links to published materials. This helps programmers connect the dots and feel more satisfied by seeing the end results of their efforts. This also gives

them insights on registry's progress and provides them a broader perspective on their contribution to the registry's evolution. Furthermore, it's recommended that programmers also seek out published materials online related to their registries. I was personally amazed to see the registry data I worked on featured in published journals; with the statistics we provided beautifully incorporated into the papers. It was a very fulfilling experience.

RECOMMENDATION 5 – Study leads should communicate the overall goals and milestones of registry projects to the programmers, share links to published materials, and encourage them to seek out published results related to their work. This approach helps programmers connect their efforts to the larger registry progress, providing them with a sense of purpose and satisfaction as they see the impact of their contributions in published journals and other outputs.

CONCLUSION

The adoption of above methodologies for real world data reporting through registries has shown significant improvement in turnaround time of the analysis and better hold on data. It has drastically reduced the change requests and has helped build better relationships between the programmers, study leads and reviewers. These approaches have significantly improved the overall data quality and health of the program.
Working on Real World Data provides ample opportunities for programmers to explore and apply standards. Rather than perceiving challenges as roadblocks, they should be seen as opportunities to develop roadmap of innovative solutions.

"It is not the strongest of the species that survive, nor the most intelligent, but the one most responsive to change." *— Charles Darwin*

In this rapidly evolving environment, those who embrace change and remain flexible are the ones who thrive and continue to progress. Be curious, be agile and adapt.

REFERENCES

https://www.cdisc.org/kb/articles/considerations-using-cdisc-standards-observational-studies

https://www.ncbi.nlm.nih.gov/books/NBK493611/

ACKNOWLEDGMENTS

The authors gratefully acknowledge Kelly Cook and James Gunter for their support.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:
Author Name: Neha Srivastava
Company: Fortrea Inc.
Email: neha.srivastava@fortrea.com

Author Name: Sivasankar Konda
Company: Fortrea Inc.
Email: siva.konda@fortrea.com