Paper DH08

If data is the new oil, how do you make the most of your reserves

Simon Luo, Hoffmann-La Roche Ltd, Mississauga, Canada
Ashley Mao, Hoffmann-La Roche Ltd, Mississauga, Canada

Abstract

Since British mathematician Clive Humby stated, “Data is the new oil”, in 2006, many companies and industries have been leveraging data to generate business insights. At Roche, we have collected a considerable amount of data from trials for various molecules and indications. In the past, upon the completion of studies, most trial data were put in repositories; as if treasures sank to the bottom of the ocean. However, in recent years, we have taken significant steps to revive and reuse these data for secondary purposes following “Findable, Accessible, Interoperable, and Reusable” (FAIR) principle; data mart is one of these efforts. In this paper, we will define what a data mart is, how we curate, harmonize, integrate and anonymize data using Roche and CDISC standards, and how data mart is utilized to drive clinical research through a Shiny app. We will also discuss the challenges, important considerations and lessons learned from an integrator perspective.

Keywords: Data Mart, Integration, Harmonization, Anonymization, FAIR, Data sharing

Introduction

As petroleum products contribute to the 20th century economy, we are seeing a new driver – data, behind almost every industry in the 21st century. Alongside the improved computing power nowadays, data is becoming more and more important and valuable to a company for business insight generation and decision making.

In this paper, we discuss how data is used to fuel the economic engine in this new digital world, and how Roche uses our data reserve - data mart assets, to help drug development. We explain what a data mart is, how to create one and the application of data mart using our Lupus Data mart as an example.

Is Data Really the New Oil? - We Believe It Is

We are facing a drastically fast changing world, politically, economically and environmentally. Think back to what we have experienced since the COVID-19 pandemic hit us in 2019. These changes require companies to think and act differently, as old business models and traditional solutions may not work anymore. In order to survive the future, industries and companies need revolution. Like any previous industry revolution, believe it or not, we need data as the new “oil” to drive the engine.

Companies such as Google, Amazon, Facebook, as well as other more conventional businesses and industries such as agriculture and retail stores, are collecting and utilizing data to improve their products/service and discover new target customers, to automate business processes and to help people be more connected.

Data plays the same role in the pharmaceutical industry. To meet our patient and business needs, we need to develop more drugs faster with a lower cost. To achieve this, we have seen many emerging innovative methods and ideas such as smart devices for data collection, synthetic treatment arm utilizing real world data, adaptive trial design, etc. All these have changed the way we develop drugs.

What is Roche Doing for Secondary Use of Data?

In most of the cases, pharmaceutical companies would collect a large amount of patient data with high quality in clinical trials. Those valuable data might be used primarily for the clinical trial reporting or are only available to the original sponsor, thus the full potential of the data is not maximized.

At Roche, we are sitting on a huge mountain of historical data that was collected from thousands of trials for various indications and molecules. In addition to conventional clinical data, we have a wide spectrum of data types including imaging, audio, eCOA, genomic and biomarker data. They sit loosely in the repositories and are untapped after the

primary use for clinical study reporting or filing support. Secondary use, on the other hand, enables data to be used in the exploratory settings other than the objectives specified in the protocols.

There has been great effort at Roche recently to curate and integrate such data to generate scientific insights to help drug development. By exploring the full potential of data, we bring great value to our patients, our colleagues and teams, and our business. With a patient-centric mindset, maximizing the secondary use of data advances early clinical research and deepens our understanding of disease, which ultimately potentiates the access to personalized medication at a reduced cost. To Roche internal team and colleagues, the secondary use of data provides a platform to share knowledge more efficiently and enables broader access to wider teams beyond the primary data resources. Finally, re-use of data allows the pooling and integration of knowledge across the organization to deliver better and timely healthcare solutions to our patients.

About Data Marts

What is data mart

Data mart is one pillar of Roche's Enhanced Data & Insights Sharing (EDIS) program to achieve the aforementioned goals. A data mart is a collection of harmonized data modalities from multiple studies and/or external data sources that are available in a single place and can be analyzed by scientists throughout the company. As of now, Roche has built over nineteen data marts and more are planned.

Each study includes different sources of clinical data and biomarker data, as well as other data types that can be linked together to form an integrated data asset. Combining multiple integrated study datasets, whether they come from internal or external studies, creates a data mart. The fundamental purpose of a data mart is to harmonize endpoints, labs, and other critical variables across studies so that scientists can compare from one study to another.

How a data mart is created

Like crude oil needs to be refined, in order to make the data comparable across studies, source data need to be curated and harmonized first according to the same standard and naming conventions. For example, lab tests, biomarker names need to be harmonized so they have the same parameter names and units. All curated data will be mapped to SDTM domains with an unique identifier that can link them.

Once the SDTM data is curated and harmonized, Integration can be performed to generate the analysis ready ADaM dataset according to the predefined analysis planned and data specification. This part is similar to the safety and/or efficacy pooling in New Drug Application (NDA) filings, but has a more complete data picture with more data points and variables of interest. Both curation and integration work can be done using SAS or R tools.

Depending on needs, the data mart team may provide an accompanying tools/user interface, such as a Shiny App Dashboard, for users to access the data and analysis. Once pooled SDTM and ADaM data and user tools are ready, the data mart can be released for the very first version. For access control and versioning, all curated and integrated data and documentation are stored in our internal data sharing platform, and all program codes are stored in Github.

Below are the life cycles and steps a typical data mart project will go through (Figure 1):

1. Data Mart Initiation
2. Data Mart Development and Release
3. Data Mart Insights Generation and Tracking
4. Promotion and Community Engagement
5. Data Mart Refresh, Extension, and Maintenance

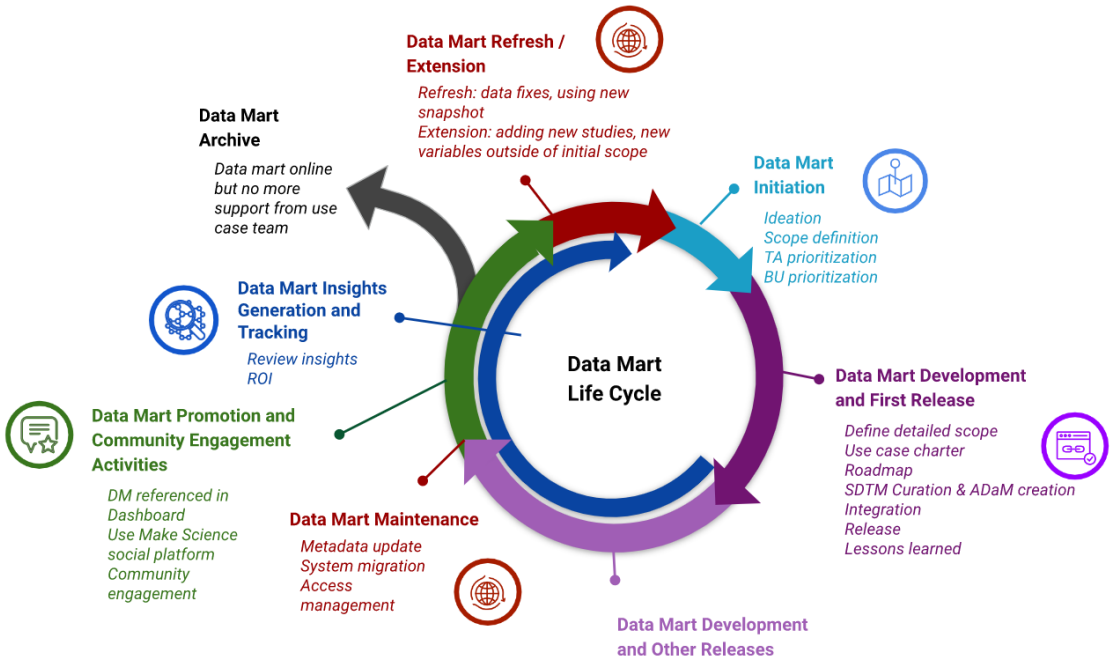

Figure 1: Data mart life cycle (Source: Roche EDIS wiki)

Challenges & Lesson Learned

Different projects may have different challenges. Based on our lupus data mart experience, we found the challenges come from the following 5 areas:

Harmonization of Efficacy Endpoint and Biomarker Data Nomenclature

The definitions of clinical efficacy endpoints across studies might be different. For example, Lupus Nephritis endpoint - Complete Renal Response definition has evolved over time from Phase II to Phase III trial, and rescue medication lists are different from protocol to protocol even for the same indication. Biomarker data nomenclature is another example, we see different lab assays were used over time. Harmonization across studies is an important and time-consuming step in preparing the source data.

Harmonization of Legacy and CDISC studies

Some data mart projects may have a mix of legacy and CDISC mandate trials or a mix of internal and external studies. The data standard, variable names, and units are different. The curators must convert those non-standard variables following our latest data standard guidelines.

Documentation

To keep track of all the discussions and decisions made during the harmonization process, and to help the end user understand the data better, documentations such as a data mart Charter, individual study protocol, Statistical Analysis Plan (SAP), eCRF, as well as pooled data specifications, user guidance, data dictionary, and exploratory analysis plan, etc. Although it is resource intensive to prepare, it becomes important for traceability and for users to understand the underlying assumptions made to the datasets.

Communications between Teams and End Users

Although the data science function plays a big role in generating the data assets, it involves multiple groups and stakeholders to complete a data mart project. A data mart team normally will consist of members from Project Management, Data Curator and Integrator, Biostatistician, Clinical Scientist, Biomarker Scientist, and other business enabling groups. Because people in the team have different backgrounds, efficient communication is critical to understand the different perspectives and requirements. Also, multiple systems and platforms were used from the development to final release. For example, the data curator and integrator have to explain CDISC terminology to the

scientist's team, and we need to work with another enabling team to get the final datasets anonymized so they can be released and shared through a platform.

Anonymization Requirements

Following data privacy protection and data sharing guidance such as GDPR, curated and integrated datasets should be anonymized. The anonymization process could cause unexpected issues for the data mart; some SDTM variables were scrambled or dropped during the process, and sometimes the linkage between the different data sources was broken in the anonymization process. This could cause integration program malfunction, and we have to think creatively to solve the issue.

Data Mart Application/User Case

With the standardized endpoint/parameters in the lupus data mart, it becomes possible to directly compare biomarkers. This significantly reduces the burden in the data processing step and puts different molecules and studies on the same scale, which further deepens our understanding of the mechanism of action and differentiation of the molecule. Scientists also found the potential of informing enrollment decisions from the data mart. With rich data sources in demographics, standards of care, biomarker, endpoint, etc, researchers were able to characterize patient populations from multiple studies. This is an example of how data mart provides scientific and clinical insights to support our clinical development programs. Lastly, data mart provides a platform to evaluate the clinical outcome and its association with different variables, enabling us to have a better understanding of the clinical response to therapeutics.

Conclusion/Future thinking

Based on our Lupus Data Mart project experience, in order to promote more secondary data use in clinical research we need to follow the FAIR principle to make our data assets more findable, accessible, interoperable, and reusable. We realized better practices could be implemented in our future data collection, SDTM mapping as well as primary data use activities. With the great effort invested in the re-use of lupus data, we start to see how the data mart helps us generate new insights and shape our decision-making in driving scientific and medical advances for lupus patients. We believe data is the long-term strategic asset for our business and our patients. We hope to see in the near future the secondary use of FAIR data will generate greater values and accelerate the clinical research, program development, and delivery of treatments to patients.

References

- Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
- Alana Harris, Roche - “How Can We Ensure Our Study Data is F.A.I.R. (Findable, Accessible, Interoperable and Reusable)?” PHUSE Connect EU 2019
- Jenet Cp, Genpro Research; Akhil Vijayan - "Overcoming Challenges of Your Integrated Summary ISS/ISE Submissions" Pharmaceutical Users Software Exchange 2022
- Fair Data Shared portal. https://fairdatashared.roche.com/

Acknowledgments

This paper and presentation has only been possible because of the reviewers and our Data Mart team members for their comments, knowledge and collaborations: Emmy Cheng, Pascal Guibord, Rhian Jacob-Moffatt, Julie Rae, Vladyslav Khudov, Nick Paszty, Li Ma, Annie Guevara, Sudhamsh Dandamraj, Hanzal Muhusin. We thank them all!

Contact Information

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

Author Name: Simon Luo
Company: Roche Canada
Email: simon.luo.sl1@roche.com

Co-Author Name: Huiyan (Ashley) Mao
Company: Roche Canada
Email: ashley.mao.am1@roche.com