Paper DS01

How can we ensure our study data is FAIR (Findable, Accessible, Interoperable and Reusable)?

Alana Harris, Roche, Welwyn Garden City, United Kingdom

ABSTRACT

The data landscape surrounding clinical trials is evolving, with data increasing in both volume and variety. A recent opportunity to work on a large-scale data pooling project, which involved the curation and integration of data across numerous studies, has had a significant effect on my future conduct as a study lead. This paper discusses some of the learnings from the project and how it has influenced my current day-to-day activities and decision making as a statistical programmer, with many aspects shifting to a longer term vision. The paper also covers our objectives as a company towards enabling the maximum use of our data, and promoting a culture of data sharing across the organisation, thereby advancing scientific insights and fuelling the vision of personalised healthcare.

INTRODUCTION

This paper will begin by giving a high-level overview of Roche's current initiatives to create a culture that enables enhanced data insights and sharing. It will then cover my involvement in one of the large-scale data pooling projects, and finish off with how it taught me the importance of our study data being FAIR (Findable, Accessible, Interoperable and Reusable).

ENHANCED DATA AND INSIGHTS SHARING AT ROCHE

WHAT IS EDIS (ENHANCED DATA & INSIGHTS SHARING)?

EDIS is a Roche-wide Program that aims to enable and promote maximum use of our data and a culture of data sharing across the organization, where the overall goal is to advance scientific insights and fuel the vision of personalised healthcare.

HOW DOES EDIS WORK?

EDIS works by prototyping ways to get reliable scientific insights more quickly and effectively through the creation of scientific use cases. The use cases may ask interesting scientific questions that, in turn, can add value to the business. These allow Roche to pinpoint where there are obstacles in finding, accessing, integrating, and sharing our data given the current state and landscape of our in-house data.

The EDIS Program then invests in projects to deliver people, processes, and IT solutions for addressing these challenges whilst also meeting the needs of the use case teams. In many instances the use cases require a large scale pooling project to be set up in order to find, access and integrate the data needed to answer the scientific questions.

CANCER IMMUNOTHERAPY DATA MART PROJECT

WHAT IS THE CIT (CANCER IMMUNOTHERAPY) DATA MART?

The CIT Data Mart is a collection of harmonized, curated, pooled and integrated data from across a number of Roche trials, including clinical and phenotype data.

The process of harmonization refers to aligning different datasets to a common standard. This is carried out by the data curation team, who can then easily pool studies together and create 'ready for analysis' data which can be picked up for use by the integration team. In the case of the CIT Data Mart these 'ready for analysis' datasets were in SDTM format.

The process of integration refers to both deriving endpoints and other necessary analysis variables, as well as

merging the clinical and genotype data together. For the CIT Data Mart, the SDTM datasets were used to create ADaM datasets which were then collated with biomarker data in an integration R script.

WHY WAS THE CIT DATA MART CREATED?

The overall aim of the data mart, driven by an EDIS use case, was to establish and scale a set of meaningful and FAIR (Findable, Accessible, Interoperable and Reusable) CIT data assets, which will enable personalized healthcare capabilities to maximize patient benefit and minimize risk.

HOW WAS THE CIT DATA MART CREATED?

After collating the locations of the SDTM datasets or each trial, a mapping tool was used to up-version each SDTM domain to a single version, SDTM v3.1.2. R programming was also used to harmonize any inconsistencies between studies. Once this step was completed and all the studies were pooled together by SDTM domain, the analysis programing could begin. Since most of the endpoints were covered in the CDISC ADaM datasets, we used the Roche standard ADaM macros and specifications, then made any necessary study or CIT specific updates.

HOW DID THE WORK COMPARE TO TYPICAL "STATISTICAL PROGRAMMIMG STUDY WORK"?

Although the main part of my role in the team was ADaM programming, it was quite different to my experience of working on a specific study. The requirements were more fluid, and able to evolve with the needs of the business. This led to the requests continually changing as the requirements from the stakeholders grew.

Given the broad, across-study nature of pooling work, it was very difficult to have in-depth knowledge of each study without being a member of each formal study team. This meant that tasks which would usually be simple when dedicated to a single study became more challenging, such as answering data collection queries (e.g. What data was collected? What domain is it in?).

WHAT CHALLENGES DID WE FACE?

Throughout the project we faced a number of difficulties. One in particular which caused delays was accessing the study data and the accompanying study documents. Most areas were restricted, including documentation such as dataset specifications. The differences in controlled terminology across the trials also caused issues. The task was not as simple as running our Roche ADaM dataset macros, as we still had to do some processing to account for differences in terminologies across the studies. Every requested endpoint and variable had to be researched and the derivation modified for each trial (depending on where the data was mapped and whether it existed). As expected there were also data anomalies, such as; death date discrepancies in AE/ DS/ DM, issues in date formats, and variables not collected/ not populated.

WHAT DID WE LEARN FOR FUTURE DATA MARTS?

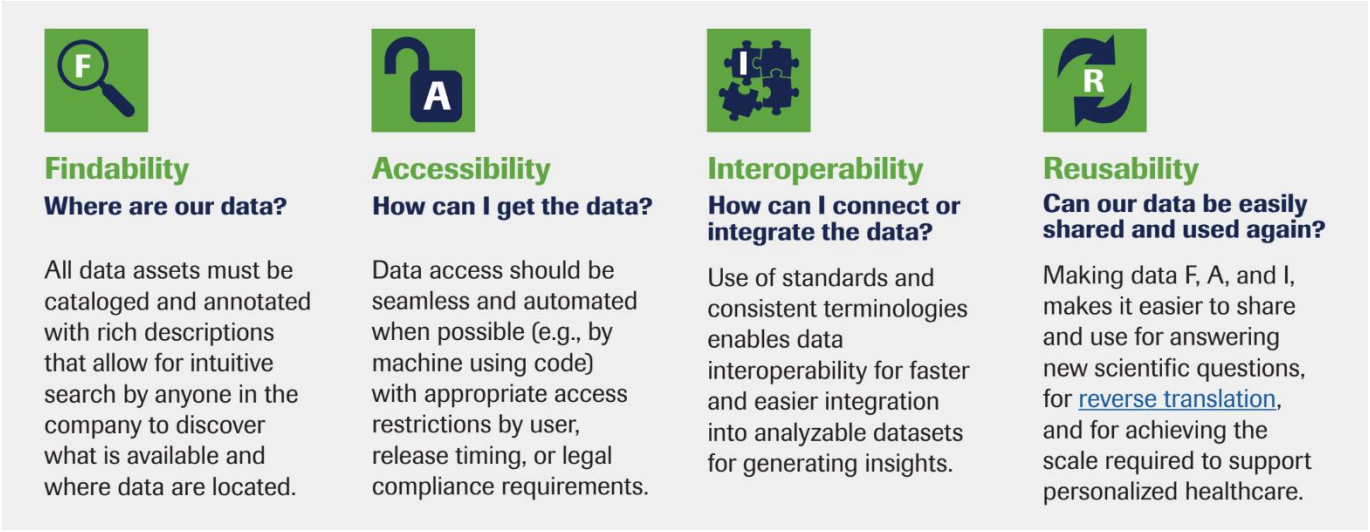
The work on this project greatly helped the team gain an awareness of ADaM standards, as well as the limitations of these when dealing with source data of an older standard. In general, the earlier the stakeholder discussions start, the better. Scope control was a real issue on the project when we had such an array of source data to use. Acquiring up to date and accurate study documents, such as; the protocol, study analysis plan, SDTM specifications and dataset specifications relating to the study's most recent Clinical Study Report, greatly helped the curation and integration process. If all our study data and documentation was FAIR, the CIT Data Mart task would have been much simpler.

FAIR PRINCIPLES

WHAT IS FAIR AND WHERE DID IT COME FROM?

The term FAIR originated from a workshop which took place in 2014 with attendees from academia, industry, funding agencies, and scholarly publishers. The group put together the FAIR Principles to describe a well-organized state of data, that is ready to share and be widely used to generate scientific insights, and allow for more informed clinical decisions. The first formal publication[1] of the FAIR principles was released in 2016, with "the intent that these [Principles] may act as a guideline for those wishing to enhance the reusability of their data.".

The following image from the Roche FAIR Data Playbook outlines what each part of FAIR relates to:



HOW DID WORKING ON THE CIT DATA MART PROJECT MAKE ME ALIGN TO THE FAIR PRINCIPLES?

Through experiencing firsthand how difficult it can be to find and access study data with up to date documentation, it really made me question whether the single studies I am currently working on are ready to be shared. A few of the hurdles we faced on the CIT Data Mart could have been easily avoided if teams had done the following, which are now the principles I always try to adhere to as a programmer on a study:

- Update study documents regularly and keep study information repositories up to date
- Store study documents in locations which are not restricted
- Try to be consistent at molecule/ indication level
- Question why the study is deemed different'
- Think about downstream impact of taking a non-standard approach

WHAT MORE COULD WE DO TO ENSURE OUR STUDY DATA IS FAIR?

Roche are very much trying to challenge the outdated mindset that the usefulness of study data ends with the conclusion of the trial. The following are some suggestions on how to get your data ready for sharing:

- Select the appropriate data standards and terminologies in the beginning
- Capture and catalog important metadata (information about data)
- Store documentation associated with the data alongside the data itself (or clearly document where to find the documents linked to a particular version of the data)
- Ensure that study documents like dataset specifications accurately describe the dataset and derivations used
- Document the workflow and processes (outlining key decisions)
- Use version control and include access to historical content (if needed)
- Comply with global data standards and terminologies (document clearly if this is not the case)

CONCLUSION

Overall I think with data increasing in volume and variety, it's crucial to be aware of the FAIR Principles, as well as how to actively practice them in our day-to-day work. We in the Pharmaceutical Industry are in a unique position with the huge amount of data we possess, however without starting to abide to these guidelines, it will become exponentially more difficult to get the full value out of the data assets we obtain and create.

REFERENCES

1. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 2016 3:160018 (2016).

ACKNOWLEGMENTS

FAIR Data Playbook (Published by F. Hoffmann-La Roche Ltd, Group Communications)

CONTACT INFORMATION

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

Author Name: Alana Harris
Company: Roche Products Limited
Address: Hexagon Place, Falcon Way, Shire Park, Welwyn Garden City
City / Postcode: Hertfordshire, AL7 1TW
Work Phone: 07823 421197
Email: alana.harris@roche.com