Paper DH09

Database Lock is Only the Beginning: The Long, Winding Journey of Secondary Data Reuse at Roche

Tamara Senior, Roche, Welwyn Garden City, United Kingdom

ABSTRACT

Traditionally in the industry, once a database is locked and study closed, it's considered the end of the road. However in recent years, some pharmaceutical companies have breathed new life into their study data by taking a novel approach and thus coining the buzzword “secondary re-use”. At Roche this work has led to critical insights, expanding disease area knowledge and furthering drug development, but it's not without its challenges. By utilising data-driven approaches and tools such as R®, Python™, LLMs, and cloud-based storage solutions, our secondary reuse of data involves gathering, curating, harmonising, and integrating data into valuable data assets. This presentation will address the challenge of harmonising data from trials conducted years ago with missingness or collected using different data standard models. And demonstrate how we integrate this clinical data with exploratory novel data types like Whole Genome Sequencing, high-dimensional exploratory biomarkers and Real World Data, among others.

INTRODUCTION TO DATA MARTS & SECONDARY USE OF DATA AT ROCHE

The Data Curation & Integration (DCI) group at Roche are Data Science Leaders delivering scalable and sustainable data assets, driving insights, and expanding disease area knowledge. DCI was established to support Roche’s Enhanced Data Insights & Sharing (EDIS) initiative which aims to repurpose existing data for new insights. This strategy has transformed Roche into more of a big data company, focusing on creating tools, systems, and changing mindsets to support this vision. Emphasising the importance of utilising collected clinical trial and external data, EDIS is driven by the FAIR* principles, ensuring data reuse and sharing both within Roche and externally.

DCI specialises in creating high-quality data assets known as Data Marts. Data marts are diverse, large-scale data assets from patient data integrated across internal trials and external consortia. A Data Mart is crucial for secondary use of data and generating insights, akin to a library with organised and curated books on a specific topic. Researchers and analysts can extract valuable insights, identify patterns, and make informed decisions. Pooling large amounts of data from multiple trials, whether internal or external, and linking clinical efficacy, safety, imaging, genomic, and other information sources can uncover new insights not evident from individual trials. Although data marts are created for disease area specific research, trials can be used outside of their main indication which can also lead to the generation of new insights.

*Findable. Accessible. Interoperable. Reusable.

CREATING DATA MARTS: CLINICAL CURATION & INTEGRATION

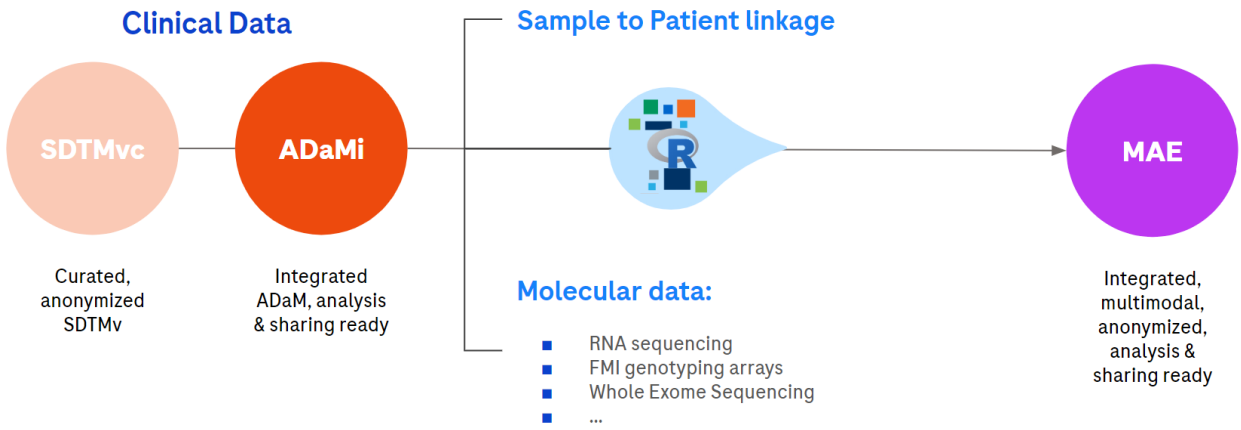

Fig 1. Data Models that make up a Data Mart

Roche Data Marts consist of 3 Data Models, a set of SDTMvc, ADaMi and MAEs (MultiAssayExperiment objects) which are harmonised and integrated before being shared (see Fig. 1). DCI collaborates with analysts to define consistent formats, follow standard naming conventions and provide clear metadata and documentation to navigate the evolving data landscape. Due to the varied data sources requiring integration, a robust harmonisation and standardisation process takes place which in DCI is called clinical curation.

CLINICAL CURATION

Whereas SDTMv is a standard format at Roche for organising and formatting data to be submitted to health authorities like the FDA, DCI delivers SDTMvc datasets, the structure and content of which are designed to enable sustained FAIRification across indications and over time, in support of secondary data usage. We maintain internal Clinical Data Curation Guidelines which are based upon the definition of SDTM in GDSR publications. The guidelines provide the information required to define what FAIR looks like for the SDTMvc datasets, allowing curation and harmonisation of data from diverse internal and external sources, with varying standards. SDTMvc datasets are created from SDTM datasets when available and this process is now highly automated, however sometimes the source data is in a non-SDTM data model. In these cases data mapping is required, and often supporting documentation and data dictionaries are used to correctly understand the data and map accordingly. As Data Marts are created using multiple studies, data harmonisation is key and is a necessary step for the downstream data pooling and analysis dataset creation.

CLINICAL INTEGRATION

Once the data has been curated to a standard format it is pooled and R is used to generate analysis ready data sets based on Roche standard specifications. Analysis ready data sets begin with a clear and unambiguous analysis plan. Teams create a data mart analysis and insights plan prior to curation work, and analysis plans and specifications are a collaborative effort continuing throughout the curation and integration process. As new questions are asked of the data, the specifications are revisited to ensure these are clearly defined. ADaMi is the DCI integrated version of the standard Analysis Data Model formatted data sets. They are pooled, harmonised datasets with derivations that are ready to answer analytical questions beyond the protocol. However an understanding of the protocol and data collection is key when pooling data as without this, the data may be misinterpreted. As part of the curation and integration process, all Adverse Event and Medical History terms are coded with the latest MedDRA version and Standardised MedDRA Queries (SMQs) and Customised Queries (CQ) are added to support signal detection and other safety analysis. All Concomitant Medications are also coded with the latest WHODrug version and dictionary derived or customised groupings are added for the analysis of specific medication classes. Harmonisation and consistency is the key to analysis across studies, but this doesn't stop with clinical data, having consistent and well represented biomarker data is essential as well.

CREATING DATA MARTS: MOLECULAR INTEGRATION

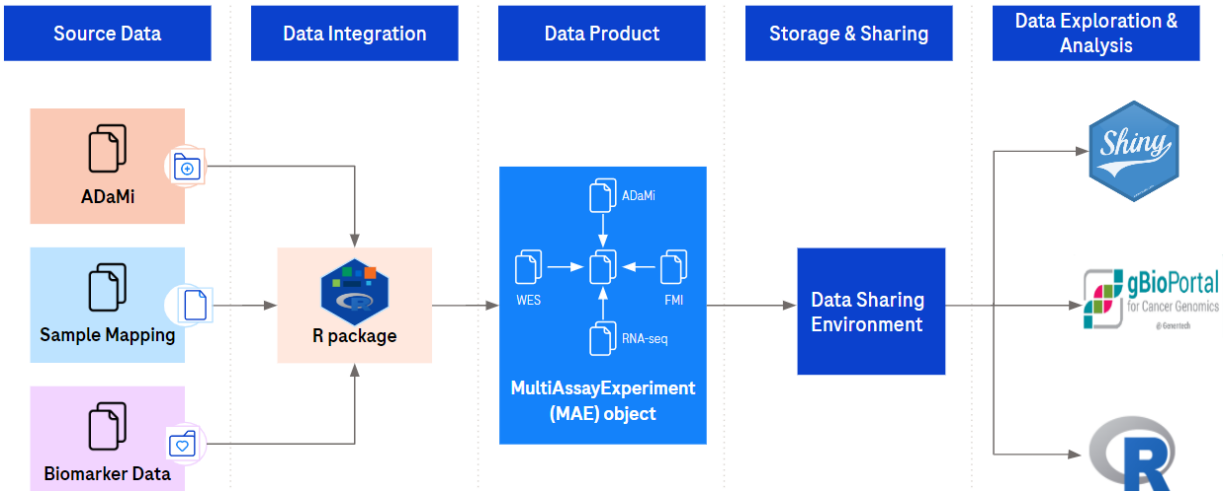

Fig 2. Molecular Integration Pipeline

MOLECULAR INTEGRATION

Having sample linkage information, allows us to integrate clinical data with other data modalities and form multimodal datasets that are important for analysis and insights. The MAEs are multi-dimensional R objects that contain clinical

(ADAMi) and biomarker (omics) data (see Fig. 2). In the context of data marts, biomarker data refers to any data generated by tests or assays which were not represented in the clinical trials data tabulations. Across our data mart portfolio we currently have 11 different high-dimensional assay biomarker types - the most predominant are gene data (Whole Genome Sequencing, Whole Exome Sequencing) and gene expression data (RNA Sequencing), but there's a very diverse group of other assays which is continually growing. MAEs are an open source data model and is the data standard for high-dimensional biomarker data. The use of a 'common data model' enables automation of the pipelines and simplifies integration with other solutions. A suite of internal R packages have been designed for creating and managing MAE objects, and delivering this format simplifies data handling for analysis as 100s of tables can be stored in a single object.

LEARNINGS

Through Data Mart creation many lessons have been learned around the optimal conditions for secondary data reuse to take place. Starting at the very beginning, if secondary use of data is considered during protocol development and every subsequent step along the way this can help to avoid issues downstream, and ensure there is an understanding of the impact of inconsistent data collection and data standards. Data standards are essential in the secondary space as much as the primary space, as having robust standards is key to automation and interoperability, and automation is key when addressing scalability and efficiency. A feedback loop with active study teams has shown to be helpful for both parties when addressing these challenges.

At the beginning of the Data Mart process which involves sourcing study data, context around the data, e.g. documentation, assay information, and study personnel, has proved just as important as the data itself. In order to correctly interpret the data and manage subsequent pooling and integration steps, having access to supporting study documentation and data dictionaries are often key. Consistency across identifiers (e.g. Sample ID) is also critical, specifically for sample to patient linkage. At the end of the Data Mart process when the data is ready to be shared, metadata about datasets, studies and molecules have proved critical to reusability and within DCI we have prioritised making this part of the entire E2E data lifecycle. In order to fulfil FAIR data standards, metadata is key for both findability and reusability of data.

IMPACT

As of 2024 there are 19 Roche Data Marts, with over 250 studies, spanning 49 indications and containing data for over one quarter of a million patients. 52 Research & Development Enabling Insights have been generated from Data Marts since 2020 and this has had a significant impact. Examples of Data Mart generated insights are identification of new indications and targets, new links between safety signals and ethnicity, new assay development and imaging algorithm development. Notable impacts include accelerated timelines, reduced cost, informing no-go decisions and new study design.

ACKNOWLEDGMENTS

The work presented in this paper is the result of the collaborative efforts of the Roche DCI Team

CONTACT INFORMATION

Your comments and questions are valued and encouraged.
Author Name: Tamara Senior
Email: tamara.senior@roche.com