Unleashing Data's Full Potential through Integrated Data Marts

Aimee Keel[1], Alexander Arefolov[1], Joseph Kochmanski[1], Lasya Sreeramagiri[2-3], Sandhya Balasubramanian[2-3]

[1] Rancho BioSciences LLC, PO Box 7208, Rancho Santa Fe, CA 92067

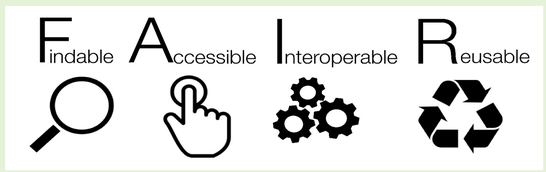
[2] Genentech, Inc., South San Francisco, CA, USA

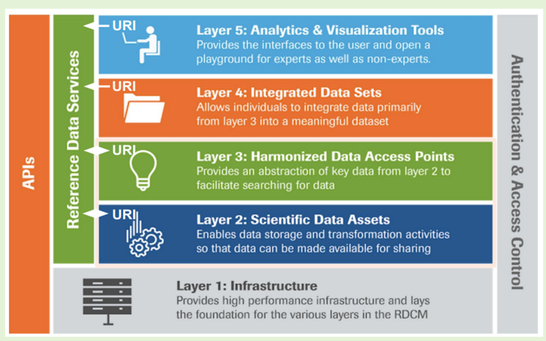
[3] F. Hoffmann-La Roche Ltd, Basel, Switzerland

Background: Giving Data a Second Life with Roche Data Marts

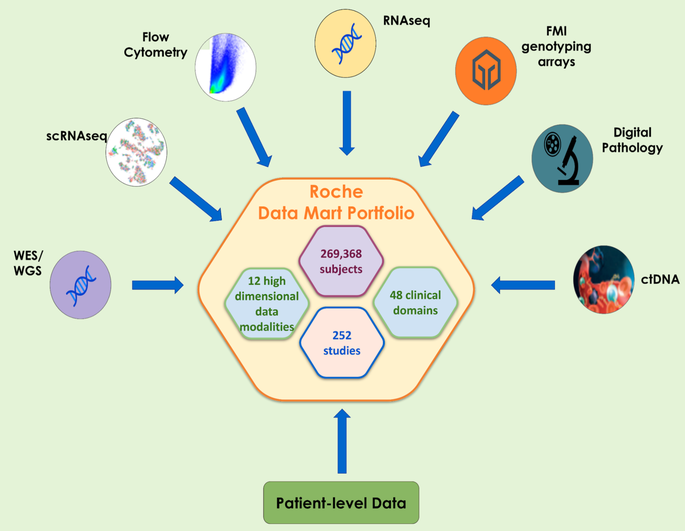
Aim: leverage Roche's large pool of existing patient data to generate insights, improve diagnostics, and drive R&D.

- To achieve that goal, the data must be **F**indable, **A**ccessible, **I**nteroperable, & **R**eusable (FAIR).
- At Roche, FAIR-ification and integration of multi-modal data is done within the **E**nhanced **D**ata and **I**nsight **S**haring (EDIS) department.
- Clinical and high-dimensional biomarker data are integrated into **M**ulti**A**ssay**E**xperiment (MAE) objects.
- Collections of MAEs for a given indication make up a single data mart.





Challenge: Integrating Diverse Data Types into Data Marts



- Standards must be established with vendor/data providers
- Evolving data models and specifications
- Lack of standardization between datasets
- Constantly evolving processing pipelines
- Data collected under very specific experimental conditions
- Inadequate capture of relevant metadata

Solution: Standardized Workflow for Diverse Data Types

Workflow for Streamlining Integration of New Data Types

0 **Engaged, early discussions with vendors to align on file format specifications: data content, output format, metadata, etc.**

1. **Biological understanding of novel data type** — Understanding of data & data exploration
2. **Evaluation of analysis potential** — Understanding of potential analysis insights and impact on clinical decision making
3. **Development of data model & specifications** — Proof of concept: development of a data model inspired by industry trends, analyst review and refinement
4. **Implementation in another therapeutic area** — Knowledge sharing with other teams, further refinement
5. **Consolidation in R package** — Data type wrappers and functions consolidated in an R package

Case Study: Integration of Global Screening Array Analyses

Global **S**creening **A**rrays (GSAs) are a faster, more cost-effective alternative to traditional **W**hole **G**enome **S**equencing (WGS). When GSA data was introduced into the data marts as a novel data type, the workflow described here was put to the test:

- Developed a **data model for raw data and each analysis product** (HLA genotyping, Polygenic Risk Scores, etc) to fit the MAE integration workflow.
- Working with data providers and upstream/downstream analysts, **refined the data model to serve all stakeholders.**
- Collected all **relevant metadata** supporting the data providence to include in the integrated object.
- Subject matter experts developed **in-depth documentation** to relay pertinent information to future integrators.
- Generated **quality control scripts** to test subject-to-sample mapping, completeness of the data, and accuracy of the data.
- **Timeline improvement:** Initial data objects were developed in approx. 3 months from the time the data could be accessed until final **S**ummarized**E**xperiments (SEs) were delivered. Now, it takes around a week to generate 168 SEs for a given data type across 28 studies for one indication/data mart.

Implementation: Each Data Mart is a Collection of Study Level MAEs Across an Indication

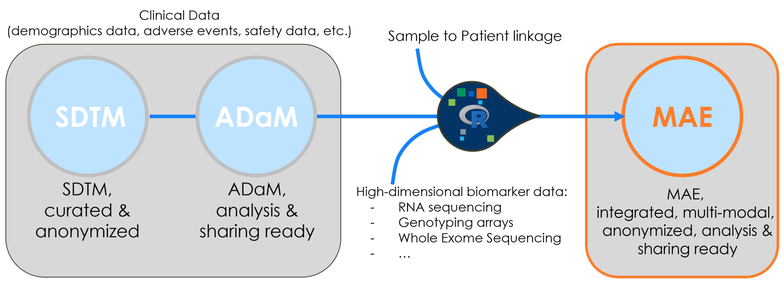
Importance of a data model for streamlining integration

- Central to the Roche internal R landscape
- Developed by the R/Bioconductor community
- Container for patient-level & high-dimensional data modalities
- In-build systematic, consistent metadata
- Standardization by design
- Consistent input for downstream storage & analyses platforms

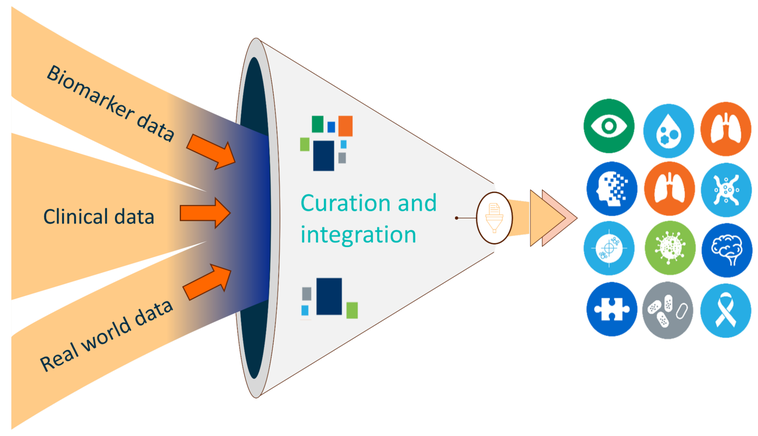
⇒ MultiAssayExperiment as a unique, self-sufficient data model

Data Marts are Delivered as 3 Data Models

Each data mart consists of a set of SDTMs, ADaMs and MAEs



The MAEs (MultiAssayExperiment objects) are multi-dimensional R objects that contain clinical (ADaM) and biomarker (omics) data. They are anonymized and available for sharing.



Lessons Learned: Overcoming Challenges of Multi-Modal Integration

The only sustainable way to manage the volume of data encompassed in the data marts is to harmonize inputs, standardize deliverables, and automate workflows across the entire end-to-end engine.

Recognizing roadblocks at the very beginning of data curation and integration can **help avoid errors**, **improve secondary data re-use**, and **enhance workflow automation**.

Potential Roadblocks:

- Non-unique sample identifiers
- Unexpected changes to upstream data outputs
- Storage size concerns
- Subtle differences between data models can impact ability to automate
- Deprecation of old data models
- Communication between multiple teams across the organization

Future Directions: Where can we go from here?

As the integration workflow becomes more automated, focus is beginning to shift to **development of visualization and analysis tools** to make the data even more accessible. **Additional data marts** will be released for high priority indications including **additional data types**. Further **infrastructure development** is also underway to solve lasting issues with data storage and retrieval. Finally, end user **support, training and tutorials** are onging to ensure the data can be used to its fullest.

Impact: Doing Now What Patients Need Next

The Roche organization's entire data landscape is working to become FAIR. Since 2020, **20 Roche EDIS data marts** have been deployed. The FAIR data included in these data marts has enabled **54 key R&D insights, allowing for faster access to data, efficient streamlining of the end-to-end engine, and cost-savings for the company**. A **single insight can lead to numerous impacts for the company**, including risk-reduction, patient safety, informed trial design, accelerated timelines, new commercial opportunities, scientific innovation, and increased revenue/savings. Additionally, the end-to-end engine has resulted in **significant critical path reduction** by shifting the focus from study level to indication level data. Many processes are suitable for **automation, resulting in a drastic decrease in resourcing needs**. Given these statistics, it is estimated that the data marts have **saved millions** for the company's end-to-end R&D engine.

FAIR data's impact for

- ✓ **Patients**: Making the most out of the patient data that we are entrusted with.
- ✓ **Teams Across the Roche Organization**: Data are more standardized and accessible, and shared in a FAIR way, generating insights far beyond what any one study team could generate.
- ✓ **Business**: Standardized data enables **faster insights**. Access to **consistent** data across indications means that R&D decisions can be made quicker and more accurately, saving money on development and generating revenue faster.

Acknowledgements

Data Curation and Integration Team, Product Development and Data Sciences (PDD), Rancho BioSciences Collaborators