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

Alana Harris



Presentation Agenda

- What is EDIS (Enhanced Data and Insights Sharing) at Roche and how does it work?
- What is the CIT Data Mart and why/ how was it created?
- What is FAIR (Findable, Accessible, Interoperable and Reusable) and what can we do to ensure our study data is FAIR?

Introduction

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

How does EDIS work?

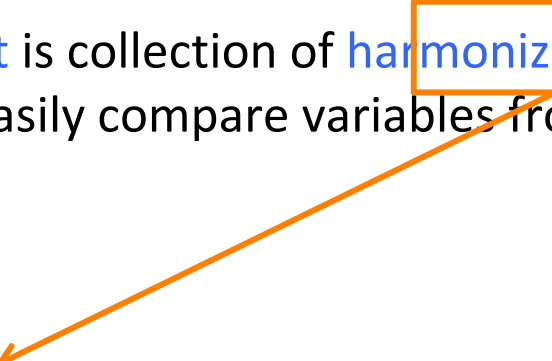
- EDIS works by prototyping ways to get reliable scientific insights more quickly and effectively through scientific Use Cases
- A Use Case will ask an interesting scientific question that can be scaled up and support data-related projects that solve longer-term business needs
- The EDIS Program invests in projects to deliver people, processes, and IT solutions for addressing these challenges whilst also meeting the needs of the Use Case teams

What is the Cancer Immunotherapy Data Mart?

- The term data mart describes a **repository of summarized data** collected for analysis on a specific section of an organization.
- The **CIT Data Mart** is collection of **harmonized, curated, pooled** and **integrated** data which allows stakeholders to easily compare variables from one study to another.

What is the Cancer Immunotherapy Data Mart?

- The term data mart describes a **repository of summarized data** collected for analysis on a specific section of an organization.
- The **CIT Data Mart** is collection of **harmonized, curated, pooled** and **integrated** data which allows stakeholders to easily compare variables from one study to another.



The task of aligning data to a common set of data standards

An orange arrow points from the word 'harmonized' in the second bullet point to the text box below.

What is the Cancer Immunotherapy Data Mart?

- The term data mart describes a **repository of summarized data** collected for analysis on a specific section of an organization.
- The **CIT Data Mart** is collection of **harmonized, curated, pooled** and **integrated** data which allows stakeholders to easily compare variables from one study to another.

The task of aligning data to a common set of data standards

Process of turning independently created data sources into unified ready for analysis data sets that conform to Roche standards

What is the Cancer Immunotherapy Data Mart?

- The term data mart describes a **repository of summarized data** collected for analysis on a specific section of an organization.
- The **CIT Data Mart** is collection of **harmonized, curated, pooled** and **integrated** data which allows stakeholders to easily compare variables from one study to another.

The task of aligning data to a common set of data standards

Process of turning independently created data sources into unified ready for analysis data sets that conform to Roche standards

Combining harmonized data from distinct groups of samples or patients

What is the Cancer Immunotherapy Data Mart?

- The term data mart describes a **repository of summarized data** collected for analysis on a specific section of an organization.
- The **CIT Data Mart** is collection of **harmonized, curated, pooled** and **integrated** data which allows stakeholders to easily compare variables from one study to another.

The task of aligning data to a common set of data standards

Process of turning independently created data sources into unified ready for analysis data sets that conform to Roche standards

Combining harmonized data from distinct groups of samples or patients

Process of creating endpoints and merging clinical data with genotype data

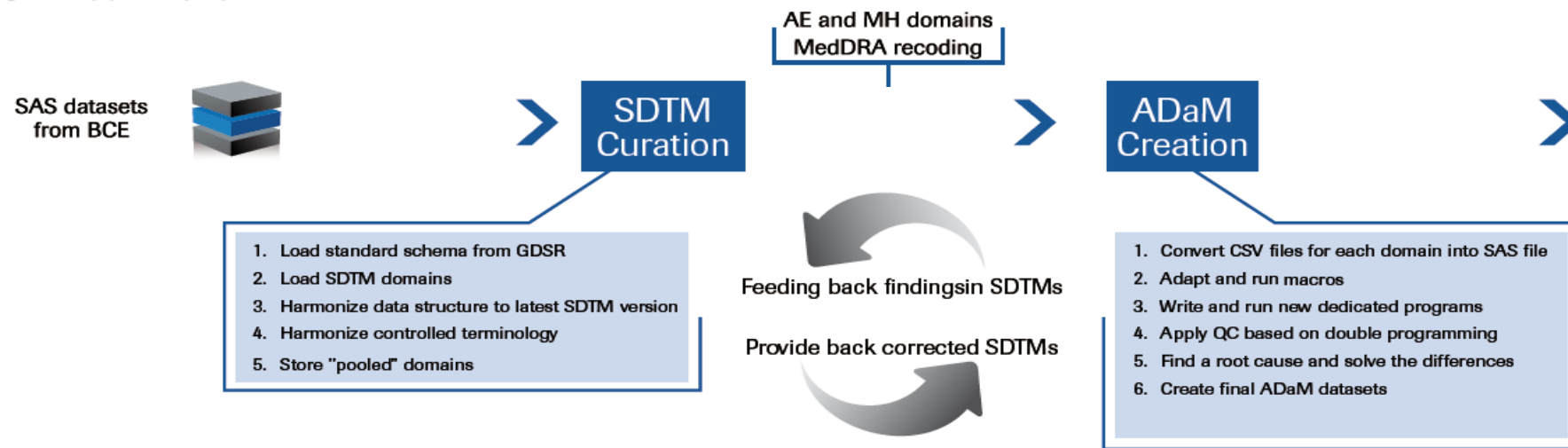
Why was the CIT Data Mart created?

- The CIT (Cancer Immunotherapy) Data Mart aimed to meet the requirements of one of the EDIS Use Cases
 - Identify patient-level genetic or environmental drivers of toxicity associated with CIT:
 1. Characterize the patient-level drivers that predict the risk and increase the understanding of underlying biology
 2. Develop predictive model to identify patient sub-groups at risk that informs clinical decision making and management of toxicity
- Establish and scale a set of meaningful and accessible CIT data assets that could enable personalised healthcare capabilities to maximize patient benefit and minimize risk

How was the CIT Data Mart created?

Curation & Integration Pipeline

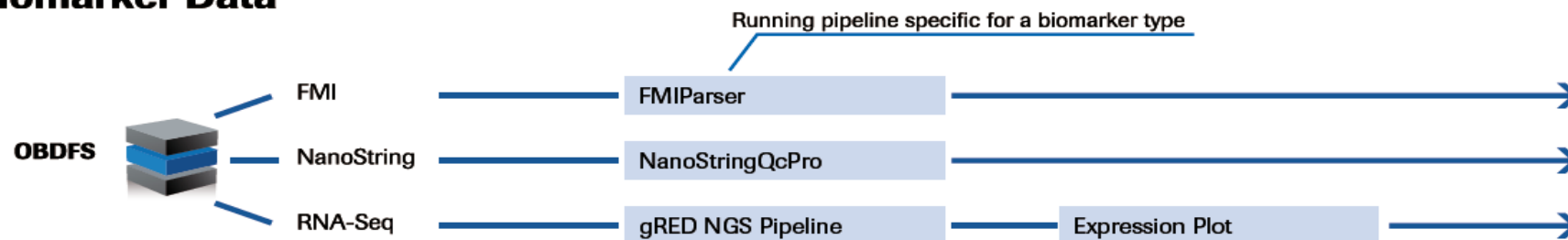
Clinical Data



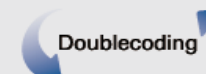
Sample Mapping Files (SMF) are produced to enable linkage of clinical and biomarker data



Biomarker Data



CLINICAL STUDY	CLINICAL SITE	CLINICAL SUBJECT	CLINICAL EVENT	CLINICAL DOMAIN	CLINICAL VARIABLE	CLINICAL VALUE	BIOMARKER STUDY	BIOMARKER SITE	BIOMARKER SUBJECT	BIOMARKER EVENT	BIOMARKER DOMAIN	BIOMARKER VARIABLE	BIOMARKER VALUE
CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001
CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001
CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001
CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001



CLINICAL STUDY	CLINICAL SITE	CLINICAL SUBJECT	CLINICAL EVENT	CLINICAL DOMAIN	CLINICAL VARIABLE	CLINICAL VALUE	BIOMARKER STUDY	BIOMARKER SITE	BIOMARKER SUBJECT	BIOMARKER EVENT	BIOMARKER DOMAIN	BIOMARKER VARIABLE	BIOMARKER VALUE
CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001
CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001
CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001
CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001

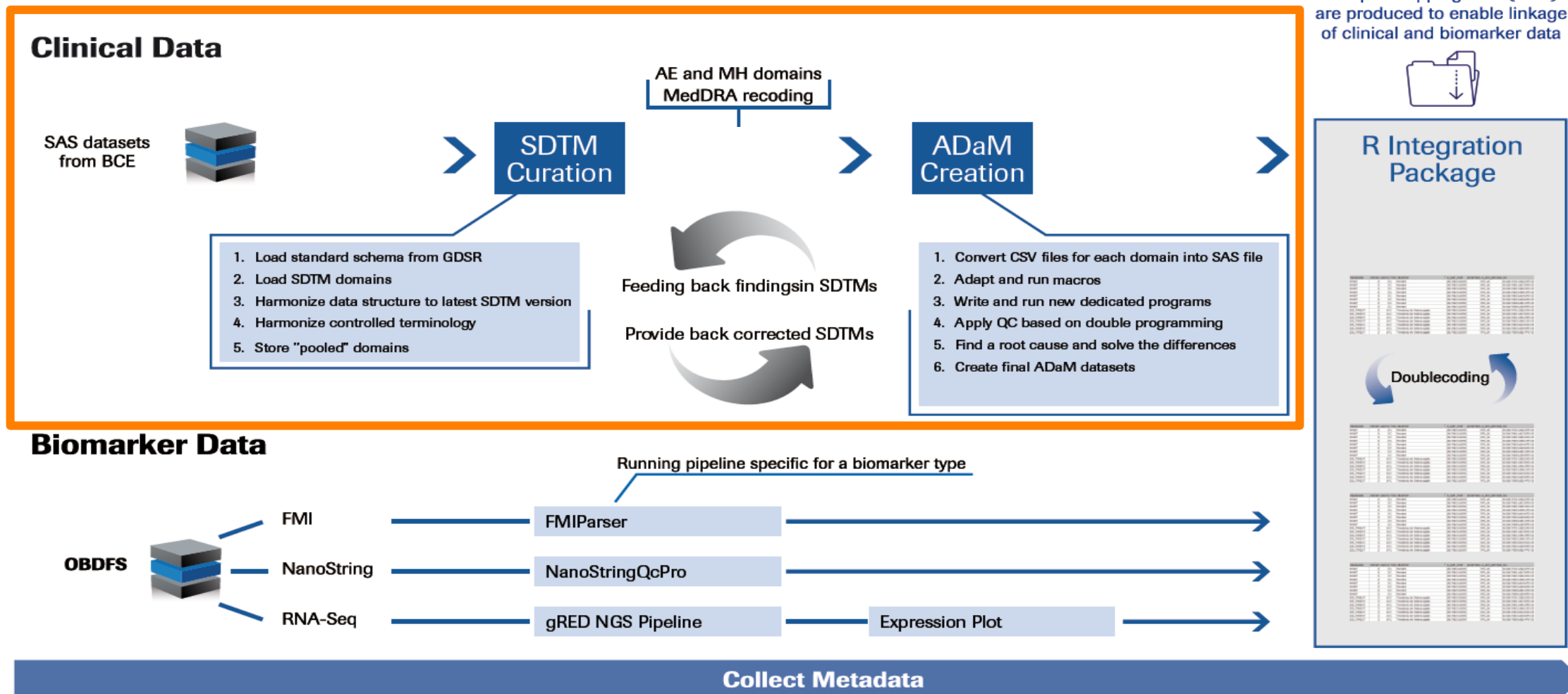
CLINICAL STUDY	CLINICAL SITE	CLINICAL SUBJECT	CLINICAL EVENT	CLINICAL DOMAIN	CLINICAL VARIABLE	CLINICAL VALUE	BIOMARKER STUDY	BIOMARKER SITE	BIOMARKER SUBJECT	BIOMARKER EVENT	BIOMARKER DOMAIN	BIOMARKER VARIABLE	BIOMARKER VALUE
CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001
CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001
CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001
CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001

CLINICAL STUDY	CLINICAL SITE	CLINICAL SUBJECT	CLINICAL EVENT	CLINICAL DOMAIN	CLINICAL VARIABLE	CLINICAL VALUE	BIOMARKER STUDY	BIOMARKER SITE	BIOMARKER SUBJECT	BIOMARKER EVENT	BIOMARKER DOMAIN	BIOMARKER VARIABLE	BIOMARKER VALUE
CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001
CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001
CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001
CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001	CTC001

Collect Metadata

How was the CIT Data Mart created?

Curation & Integration Pipeline



How was the CIT Data Mart created?

1. Located SDTM datasets for each trial
2. A mapping tool was used to up-version the SDTM data to a common standard (R programming also used)
3. Study data then pooled by SDTM domain
4. Produced the following ADaM datasets from the pooled SDTM:
 - ADAE, ADCM, ADLB, ADMH, ADRS, ADSL, ADTTE, ADSAFTTE, ADSUB, ADTR, ADVS, ADZB, ADEX
 - Most of the endpoints were covered in the CDISC ADaM datasets so the Roche standard ADaM SAS macros and specifications were used, then any necessary study or CIT specific updates made as needed
5. ADaM datasets were then integrated with biomarker data to create the final CIT Data Mart

Outcomes/ Feedback

- The Data Mart has been shared across the company for Scientists to use in their research
- *'I must admit that the CIT Data Mart saves me, a human geneticist, months (I daresay years) of time confirming the correct phenotypic variables and aligning them across studies, and then correctly integrating it with tumor biomarker data. Thanks to the CIT Data Mart, I have at my finger tips uniformly processed RNA-sequence data, FMI mutation data, nanostring data to integrate into my genetics analysis. I don't have to hunt down the biomarker teams to find this data. I can easily look for reproducible signals across trials and indications and not struggle with issues around data formatting and data processing.'* – Roche Human Genetics Scientist

How did the work compare to typical study work?

- No formal SAP – requirements more fluid, challenging to keep up
- Scope increase - requirements from the stakeholders grew
- No in depth study knowledge
- No formal study team to ask queries to (e.g. What data was collected? What domain is it in?)



What challenges did we face?

- Getting [access to study data and documents](#)
- Differences in [controlled terminology](#)
- Every requested [endpoint and variable had to be researched](#) and the derivation modified for each trial
- Discovered [data anomalies](#)



What did we learn for future data marts?

- Awareness/ limitations of [ADaM standards](#)
 - Difficulty dealing with source data of an older standard
- The earlier the [stakeholder discussions](#) start the better!
- Importance of receiving up to date and accurate [study documents](#)
- **If our study data and documentation were FAIR, creating the CIT Data Mart would have been much simpler!**



What is FAIR (Findable, Accessible, Interoperable and Reusable)?

- FAIR originated from a workshop with attendees from academia, industry, funding agencies, and scholarly publishers
- FAIR Principles describe a **well-organized** state of data, that is **ready to share** and be widely used to generate scientific insights
- Published in 2016 - *The FAIR Guiding Principles for scientific data management and stewardship*, Wilkinson, M. D. et al.

What is FAIR (Findable, Accessible, Interoperable and Reusable)?



Findability

Where are our data?

All data assets must be cataloged and annotated with rich descriptions that allow for intuitive search by anyone in the company to discover what is available and where data are located.



Accessibility

How can I get the data?

Data access should be seamless and automated when possible (e.g., by machine using code) with appropriate access restrictions by user, release timing, or legal compliance requirements.



Interoperability

How can I connect or integrate the data?

Use of standards and consistent terminologies enables data interoperability for faster and easier integration into analyzable datasets for generating insights.



Reusability

Can our data be easily shared and used again?

Making data F, A, and I, makes it easier to share and use for answering new scientific questions, for [reverse translation](#), and for achieving the scale required to support personalized healthcare.

How has the CIT Data Mart project influenced decisions in my role?

- Change of **mindset** - our data will be used again
- Update **study documents** regularly
- Store study documents in locations which are **not restricted**
- Be **consistent** at molecule/ indication level
- Question why **our study is deemed “different”**
- Consider the downstream impacts of taking a **non-standard approach**



What more could we do to ensure our study data is FAIR?

- Select the [appropriate data standards and terminologies](#) in the beginning
 - Comply with these standards throughout the study!
- Capture and [catalog important metadata](#)
- Store documentation associated with the data alongside the data itself (or [clearly document where to find documents](#))
- Ensure that study documents [accurately describe the dataset](#) and derivations used
- Document the [workflow and processes](#)
- Use [version control](#) and include access to historical content

Concluding Comments

- Crucial to be aware of the FAIR Principles and how to practice them
- Lots of benefits to being FAIR!
- Unique position with the huge amount of data we possess
- Will become more difficult to get the full value out of the data assets we obtain and create!

Questions?



***Doing now what patients need
next***