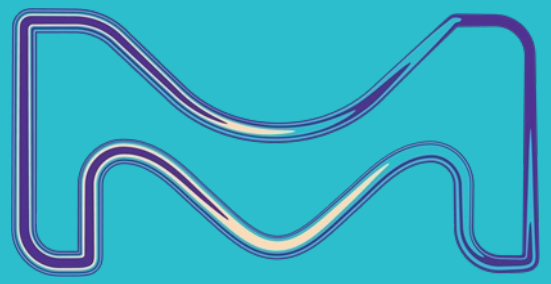
BIOANA library

SAS macro library for biomarker data quality checks

Violeta Labarta Beceiro, Samaneh Sabiani
PHUSE Heidelberg, Jun 27th, 2024



MERCK

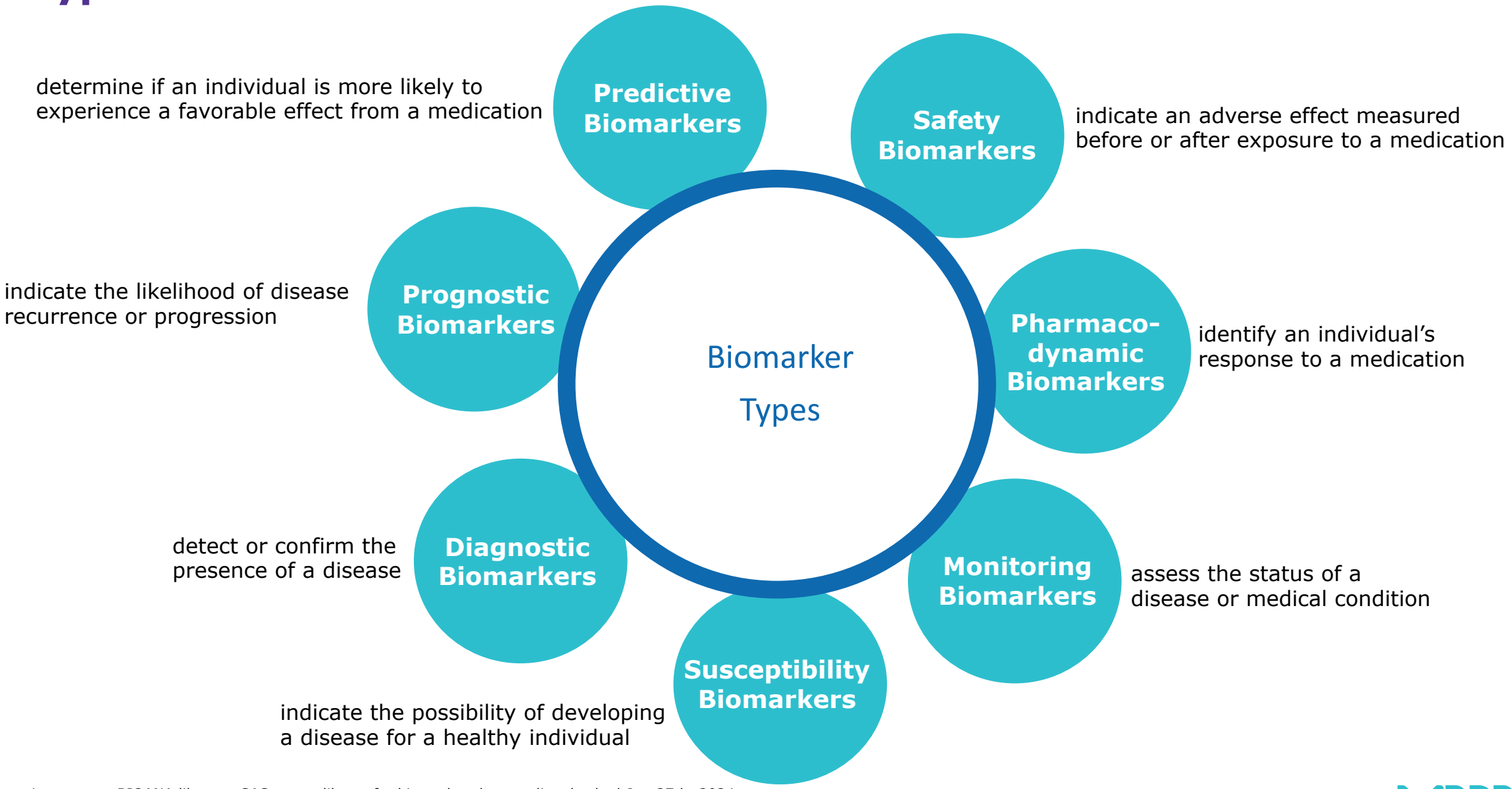
Agenda

1	Biomarker definition and utilities
2	BIOANA data QC library development background and history
3	BIOANA library: data content, reconciliation and structure checks; programming strategy
4	Data quality check example, key factors
5	Summary, Vision
6	Q&A

What is a Biomarker?

A **biomarker** is a measurable indicator of a biological or medical state or condition. This can include a wide range of biological molecules and characteristics, such as proteins, genes, hormones, cells, or physiological processes, that can be objectively measured and evaluated.

Types of Biomarker



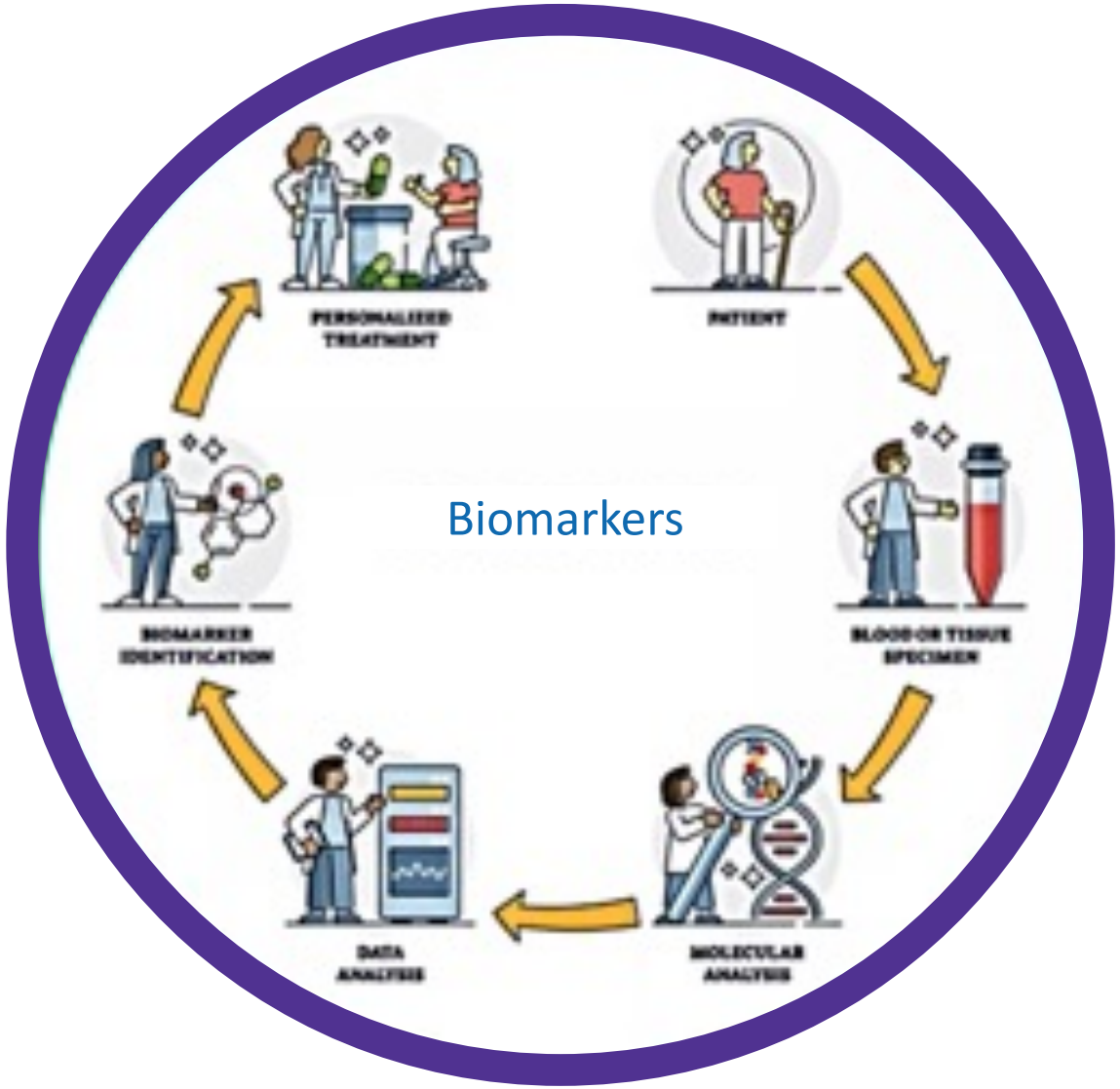
Biomarker in Precision and Personalized Medicine

Precision medicine:

Targeted Therapies based on Molecular Diagnostics

Example:

In lung cancer EGFR Mutation serves as a biomarker to identify patients who are likely to benefit from targeted therapies such as erlotinib or afatinib.



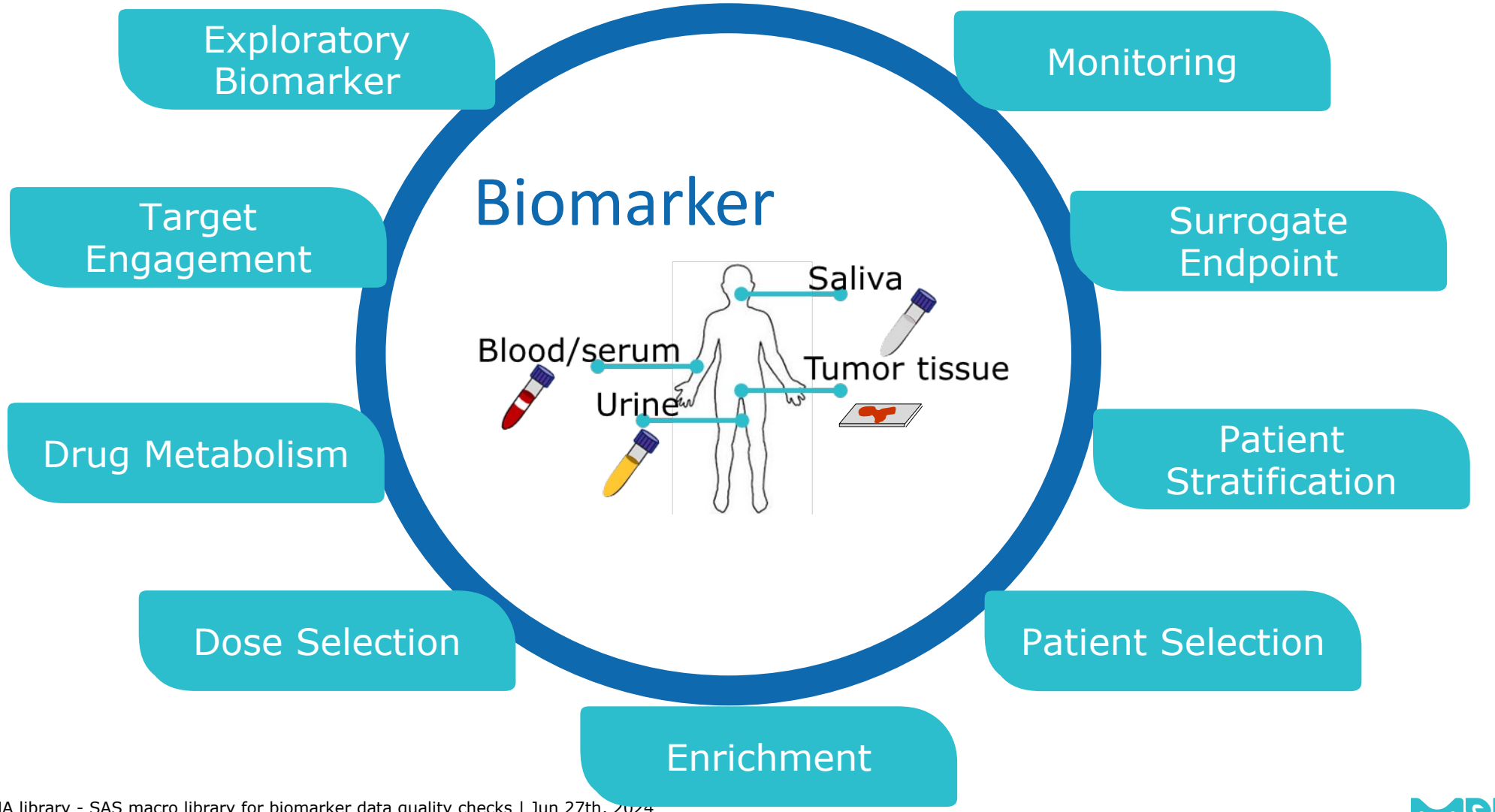
Personalized medicine:

Prevention and Treatment based on Environment, Lifestyle, and Genes

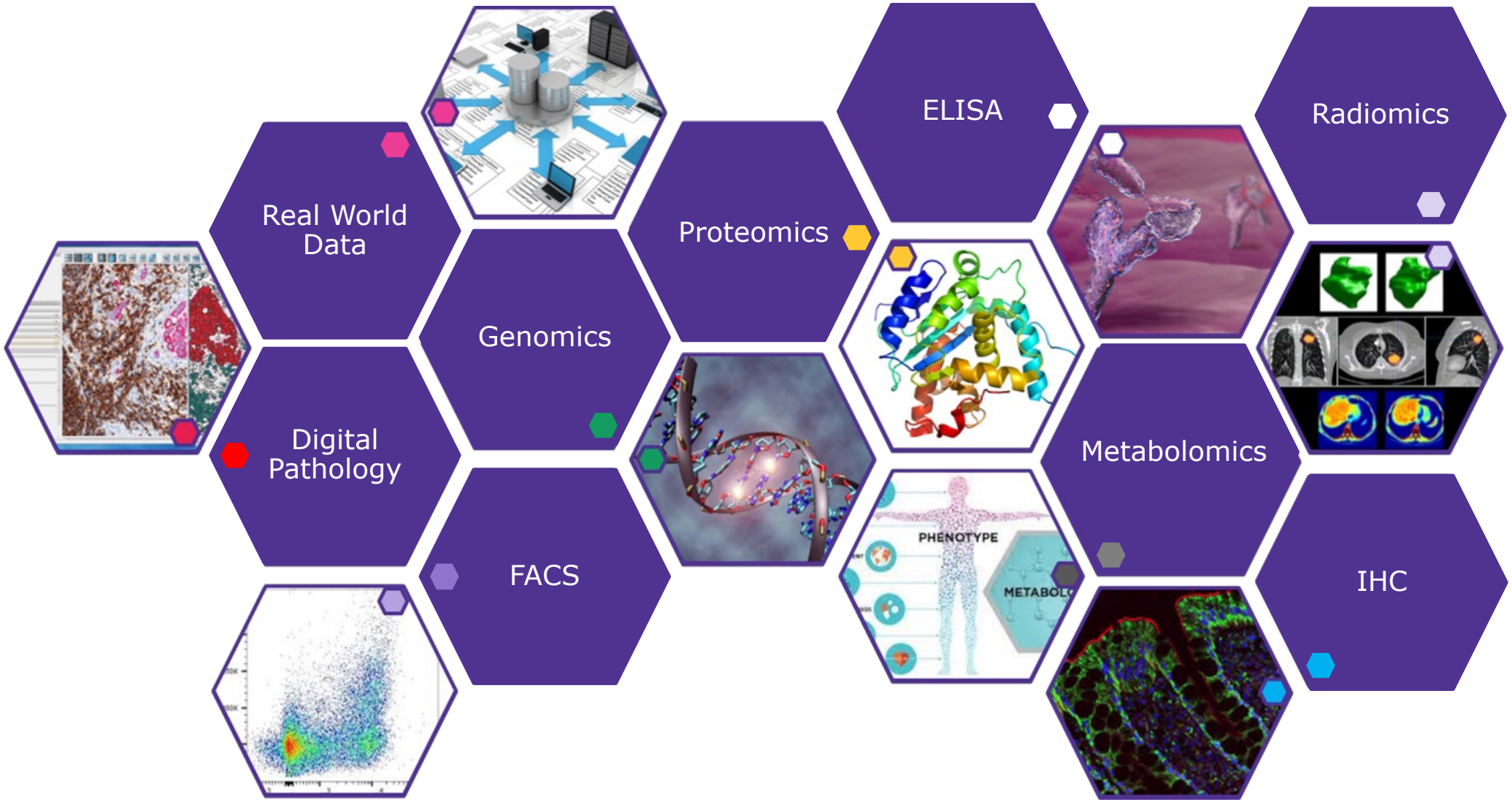
Example:

In cardiovascular disease, the activity level of the CYP2C19 enzyme serves as a biomarker for determining an individual's response to clopidogrel, a commonly prescribed antiplatelet medication.

Biomarker in Drug Development



Biomarker Identification Techniques



Biomarker Data Example

MATERIALS TESTED

Source material:
Source block:
Outside accession number:
Collection date:

DNA Sequencing data received as pdf file

MOLECULAR RESULTS

CpG Island Methylation Phenotype (CIMP) Panel Methylation Analysis:

Testing detected no methylation in any of the 3 markers tested (CIMP negative).

Methylated Markers:

METHODOLOGY: Assay is performed using either formalin-fixed, paraffin-embedded tissue blocks or frozen tissue samples. DNA extracted from formalin-fixed, paraffin-embedded tissue or frozen tissue samples is treated with bisulfite to convert methylated cytosine to uracil. PCR amplification of both unmethylated and methylated MINT1, MINT2 and MINT31 loci, and promoter sequences of p14, p16 and hMLH1 genes is performed and methylation status is assessed by pyrosequencing. The tumor is considered CIMP High, if at least 40% of markers tested show methylation, and CIMP Low if < 40% markers show methylation.

The following markers were not assessed: MINT2 *p14* *p16*

**This test was developed and its performance characteristics determined by [redacted] It has not been cleared or approved by the U.S. Food and Drug Administration. The FDA has determined that such clearance or approval is not necessary. This test is used for clinical purposes.

Entire report and diagnosis completed by: [redacted]

Biomarker Data Example

Immunoassay data structure from vendor.

Sample ID	Visit	Total protein Conc (µg/mL) *(CV%)*	Total c-MET (pg/mL) *(CV%)*	Cut-off values Y1234/Y1235 Phospho c-MET	Y1234/Y1235 Phospho c-MET (MFI - cut-off) *(CV%)*	Cut-off values Y1349 Phospho c-MET	Y1349 Phospho c-MET (MFI - cut-off) *(CV%)*	pg tMet/ mg Prt	Y1234/Y1235 pMet/ ng corrected tMet	Y1349 pMet/ ng corrected tMet	Y1234/Y1235 phospho c-Met % inhibition	Y1349 phospho c-Met % inhibition
[redacted]	Screening	1107.4 (24.2)	1167.3 (10.6)	118,6	< cut-off	88,5	< cut-off	1054,1	N.A.	N.A.	N.A.	N.A.
[redacted]	C1D17	1955.8 (8.0)	3000.1 (14.4)		< cut-off		< cut-off	1534,0	N.A.	N.A.		
[redacted]	Screening	1254.1 (5.4)	4359.0 (13.6)	118,6	< cut-off	88,5	< cut-off	3475,8	N.A.	N.A.	N.A.	N.A.
[redacted]	C1D17	450.0 (24.8)	119.4 (12.7)		< cut-off		< cut-off	265,3	N.A.	N.A.		
[redacted]	Screening	1065.4 (28.8)	9515.6 (15.9)	121,9	< cut-off	86,3	< cut-off	8931,5	N.A.	N.A.	N.A.	N.A.
[redacted]	C1D17	66.7 (15.1)	508.7 (10.6)		< cut-off		< cut-off	7626,7	N.A.	N.A.		
[redacted]	Screening	951.8 (9.2)	15437.5 (13.5)	121,9	10.56*	86,3	18.52*	16219,3	0,651	1,142	100,0	65,1
[redacted]	C1D17	741.2 (12.9)	47925.7 (12.7)		< cut-off		25.72*	64659,6	N.A.	0,398		

BIOANA Library Development
Background and History

- Non-standardized biomarker data
- Different vendors with basic DM capabilities
- Different internal DM capabilities (Multiple software & methods)
- Unreconciled sample info & unclean biomarker data

- SAS as validated programming Language for clinical data
- Entimice as GCP data environment
- Data consistency check
- Data visualization
- Reconciliation of sample info
- First run of BIOANA Library

- Data structure check
- Internal data standards for non-genetic biomarker data
- CDISC working groups for lab data findings (GF, CP, …)

- BIOANA library documentation release

- Internal data standards for genetic biomarker data

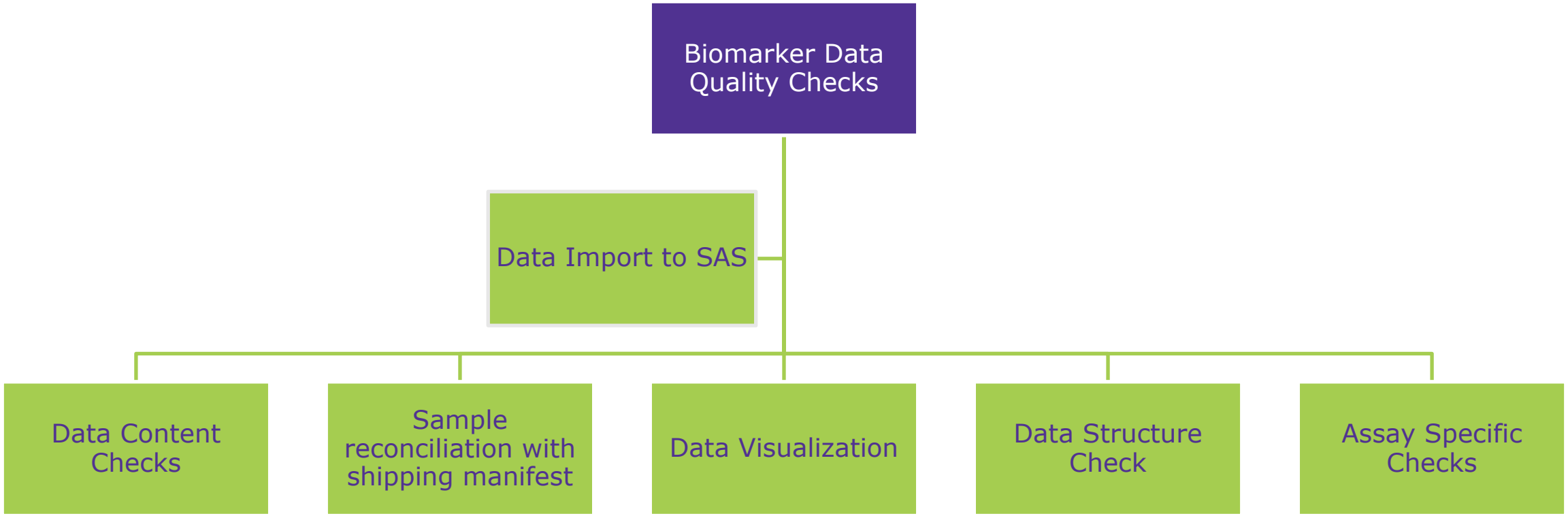
- SDTMIG 3.4 with new finding domains GF, CP, …
- SDTM from source

- Generic DTS template for non-CRF data

Programming project => Collaboration project

Programming develops based on team and multiple functions communication.

Data Quality Checks Development



BIOANA library: data content, reconciliation and structure check

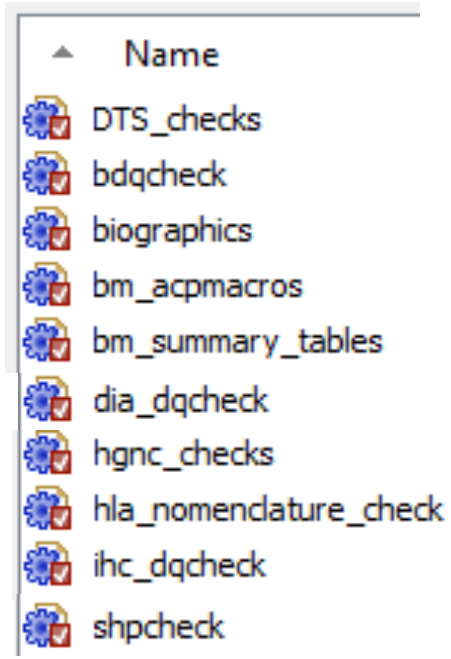
Data Content Check

Data Import to SAS

Date formats

Frequency tables

Assay specific check



Visualization

Data Structure check

Significant digits

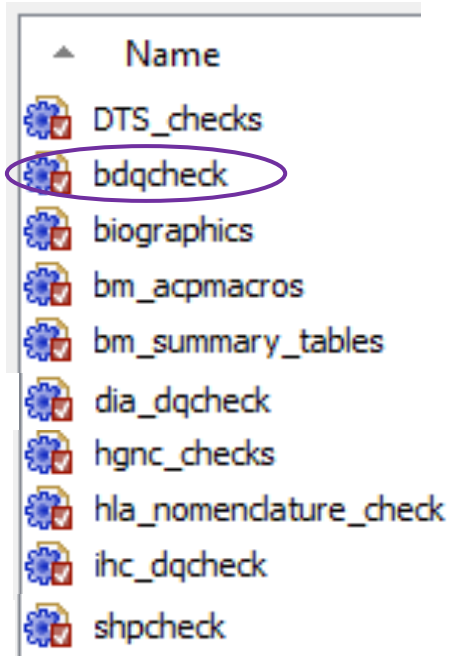
Shipping manifest quality

DTS ingestion

Library development

BDQChecks library overview

- Check 0: %file_characteristics - General Information.
- Check 1: %uniqueness - Dataset key uniqueness.
- Check 1b: %smplid1_vs_smplid2_uniqueness - Link sample ID 1 vs sample ID 2 uniqueness.
- Check 2: %shippingmanifest - Compare with Shipping Manifest.
- Check 2b: %shipping_manifest_variable - Compare with Shipping Manifest columns.
- Check 2c: %shipping_manifest_parentsmp_id - Compare with Shipping Manifest column sample ID with extension.
- Check 3: %data_integrity - Data integrity.
- Check 4: %sig_digits - Significant Digits.
- Check 5: %categories - Categorical variables.
- Check 6: %numeric - Numerical variables.
- Check 7: %twovariables - Two variables - frequency table.
- Check 8: %missing_values - Missing values.
- Check 9: %informed_consent - Informed Consent.
- Check 9b: %informed_consent_withdrawal - Check informed consent using a list of patients who withdraw consent.
- Check 10: %sum_related_values - Values Consistency: Sum of related values.
- Check 11: %all_frequencies – Frequency table for each column. (Only for exploration)
- Check 12: %check_sample_id_and_variable_x - Variable value consistency per sample.
- Check 13: %lencheck – Confirm length of values reported in a given variable are as defined.
- Report: %reception_report – QC report including all the checks already performed in the program.



Programming strategy

Flexibility and big data files challenge

Flexible data structure and terminology. Agreement with vendor is followed.

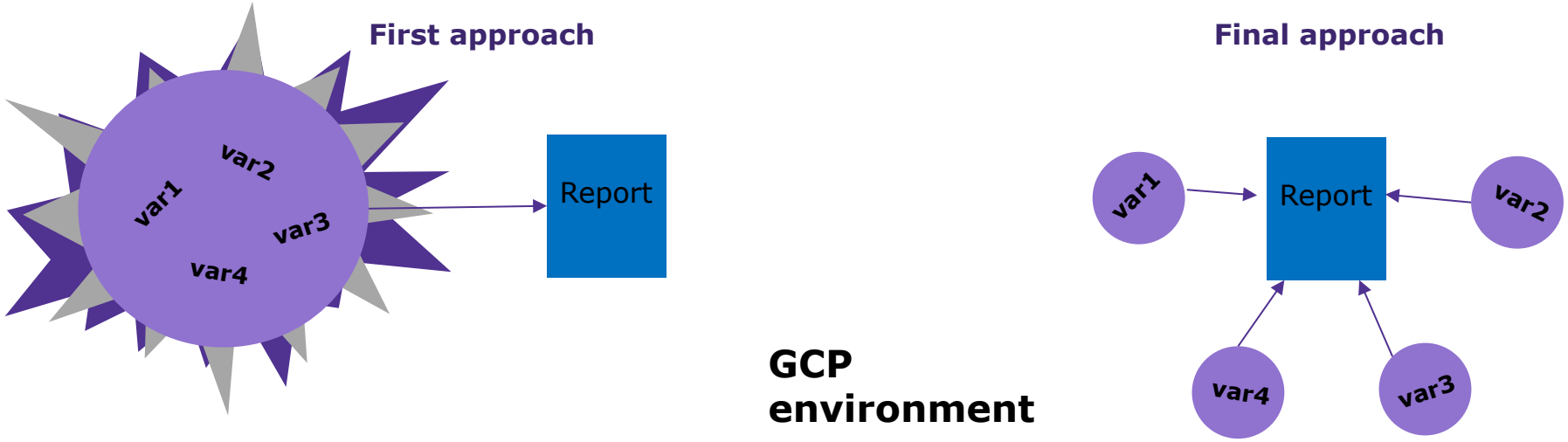
Additional user own checks are possible.

A check is done per a single variable each time, in order to reduce the data size.

The datasets are physically stored. Datasets in the work library are avoid as possible. Cleaning of the work library is performed regularly.

In the case of a big data files only few records are used initially.

Results are stored in the work library. At the end all the checks results are collected and reported.



Data Transfer Specifications

Agreement with vendor

Data Structure

Variable Name	Variable Label	Type	Length	Code list	Comment
STUDYID	Study Identifier	CHAR	20		As specified in Shipping Manifest (STUDY ID)
SUBJID	Subject Identifier	CHAR	10		As specified in Shipping Manifest (SUBJECT ID)
GFREFID	Reference ID	CHAR	30		Assay sample ID (label by vendor)
GFTESTCD	Short Name of Measurement, Test or Exam	CHAR	8	GFTESTCD	See tabs VLM and Control Terminology for details
GFTEST	Name of Measurement, Test or Examination	CHAR	40	GFTEST	See tabs VLM and Control Terminology for details
GFTSTDTL	Measurement, Test or Examination Detail	CHAR	200	GFTSTDTL	See tabs VLM and Control Terminology for details
GFCAT	Category	CHAR	200		See tab VLM for details
GFORRES	Result or Finding in Original Units	CHAR	200		To be reported only in case DROPLET COUNT CHECK PASSED. If DROPLET COUNT CHECK FAILED, column GFORRES should be reported as blank and in coumn GFREASND should be reported the reason, "DROPLET COUNT CHECK FAILED".
GFORRESU	Original Units	CHAR	50	UNIT	To be reported only if column GFORRES is populated. See tabs VLM and Control Terminology for details
GFGENREF	Genome Reference	Char	20		e.g. GRCh38.p13
GFSYM	Genomic Symbol	Char	25		See VLM
GFREASND	Reason Not Done	CHAR	200		See tab Control Terminology for details
GFNAM	Laboratory/Vendor Name	CHAR	200		Use Capital Letters
GFSPEC	Specimen Material Type	CHAR	200	GENSMP	See tabs VLM and Control Terminology for details
GFMETHOD	Method of Test or Examination	CHAR	200	METHOD	See tabs VLM and Control Terminology for details
GFRUNID	Run ID	CHAR	200		Pipeline
GFANMETH	Analysis Method	Char	200		See VLM
GFREPNUM	Repetition Number	CHAR	2		
VISIT	Visit Name	CHAR			As specified in Shipping Manifest (VISIT NAME)
GFDTC	Date/Time of Collection	CHAR	25		As specified in Shipping Manifest (COLLECTION DATE AND TIME) ISO8601
COVAL	Comments	CHAR	200		
GFMQLREV	Merck Quality Review	Char	50		Sample Processing Date ISO8601
COLSMPID	Collected Sample Identifier	CHAR	30		As specified in Shipping Manifest (ACCESSION NUMBER)
PARSMPID	Parent Sample Identifier	CHAR	30		As specified in Shipping Manifest (SAMPLE ID)

Data Transfer Specifications

Agreement with vendor

Data Quality Checks

Uniqueness:		
1	*YES*	*Check for unexpected repetitions*
Completeness:		
2	*YES*	*Check if the correct number of observations per sample are available*
3	*YES*	*Check if the expected samples are found in the data*
4	*YES*	*Check whether the data file contains all columns which are mentioned in the DTS*
Accuracy:		
5	*YES*	*Check whether a reason is given when a result is listed as blank*
6	*YES*	*Check the defined terminology according to the DTS*
7	*NO*	*Check the order of the Columns*
Consistency:		
8	*YES*	*Checking the content of the data e.g. Numerical range*
9	*YES*	*Comparison of related data e.g. --ORRES and --REASND*
Trial dependent consistency checks:		
10		*Please specify:*

Data Quality Check Report

Communication with Biomarker Team and Vendor

Data Reception Report
<Study ID>
<Data Type>

Biomarker Data Quality Check (BDQCheck)

Item	Status
General Characteristics	
Vendor	<Vendor>
Received File	<Study ID>_<Data Type>_<Vendor>_<DDMMMYYYY>_PROD.csv
Dataset Type	PRODUCTION
Transfer Type	CUMULATIVE
Dataset Use	EXPLORATORY
Additional Data Expected	YES
Biomarker Type	GENE EXPRESSION
Assay Compliance	Research Use Only
SAS dataset (after importing in SAS)	chckbmd_<Data Type>
BDQCheck Date / Time	30APR24 / 21:34
Record Number	870
Column Number	24
Subject Number (based on column SUBJID)	47
Sample Number (based on column COLSMPID)	290
Data structure according DTS - Variable Name	OK
Uniqueness	
- Unique key (Primary key: GFREFID GFTESTCD GFSYM GFTSTDTL, compare until GFTSTDTL)	OK
- Link PARSMPID vs GFREFID	OK
Completeness	

Data Reception Report
<Study Id>
<Data Type>

Biomarker Data Quality Check (BDQCheck)

Item	Status
(Sample information file)	Shipping_manifest_1.csv, Shipping_manifest_2.csv
Column selected to compare:	COLSMPID
Total of matching samples	290
Checking sample information file columns included in the data file:	
-Comparing column STUDYID in the data and column STUDY_ID in the sample information file	OK
-Comparing column SUBJID in the data and column SUBJECT_ID in the sample information file	OK
-Comparing column VISIT in the data and column VISIT_NAME in the sample information file	OK
-Difference found between column GFDTC in the data and column COLLECTION_DATE_TIME_ISO8601 in the sample information file	<gfrefid>, GFDTC= <yyyy-mm-ddThh:mm:ss>, COLLECTION_DATE_TIME_ISO8601=<yyyy-mm-ddThh:mm:ss>
	<gfrefid>, GFDTC= <yyyy-mm-ddThh:mm:ss>, COLLECTION_DATE_TIME_ISO8601=<yyyy-mm-ddThh:mm:ss>
-Comparing column PARSMPID in the data and column SAMPLE_ID in the sample information file	OK
Data integrity:	
- (1,3) records per SUBJID and GFREFID	OK
3 record(s) per sample	290 (100.0%)
Accuracy	

Data Quality Check Report

Data Reception Report
<Study Id>
<Data Type>

Biomarker Data Quality Check (BDQCheck)

Item	Status
Missing values:	
- Column GFREFID	OK
- Column GFORRES(if strip(GFREASND)='')	OK
- Column GFGENREF	OK
- Column GFREASND(if strip(GFORRES)='')	OK
- Column GFNAM	OK
- Column GFSPEC	OK
- Column GFRUNID, percent of missing per GFREFID	<gfrefid>: 100.00% of records
- Column GFREPNUM	OK
- Column COVAL(if strip(GFREASND)='' and strip(GFORRES)='')	OK
- Column GFMQLREV, percent of missing per GFREFID	<gfrefid>: 100.00% of records
Consistency	
Checking the content of categorical columns:	
- Column GFTESTCD ('GENESIG')	OK
- Column GFTEST ('Gene Signature')	OK
- Column GFTSTDTL ('NORMALIZED LEVEL' 'SCORE')	OK
- Column GFCAT ('GENE EXPRESSION PROFILING')	OK
- Column GFORRESU ('' 'copies/22uL')	OK
- Column GFSYM ('CD274' 'GUSB' '')	OK
- Column GFREASND ('DROPLET COUNT CHECK FAILED' 'INSUFFICIENT SAMPLE VOLUME' 'POOR SAMPLE QUALITY' 'FAILED QC METRICS' 'SAMPLE LOST' 'SAMPLE DESTROYED' '')	OK
- Column GFSPEC ('RNA')	OK

Data Reception Report
<Study Id>
<Data Type>

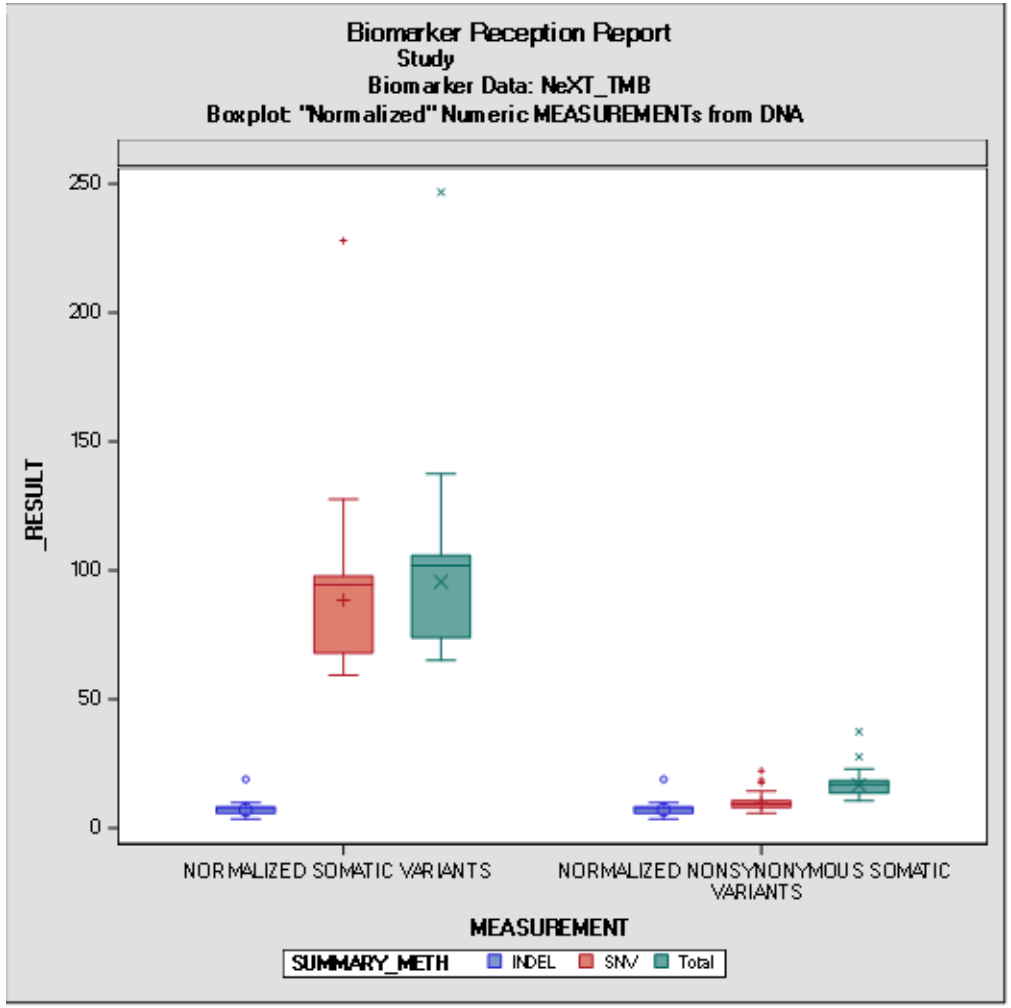
Biomarker Data Quality Check (BDQCheck)

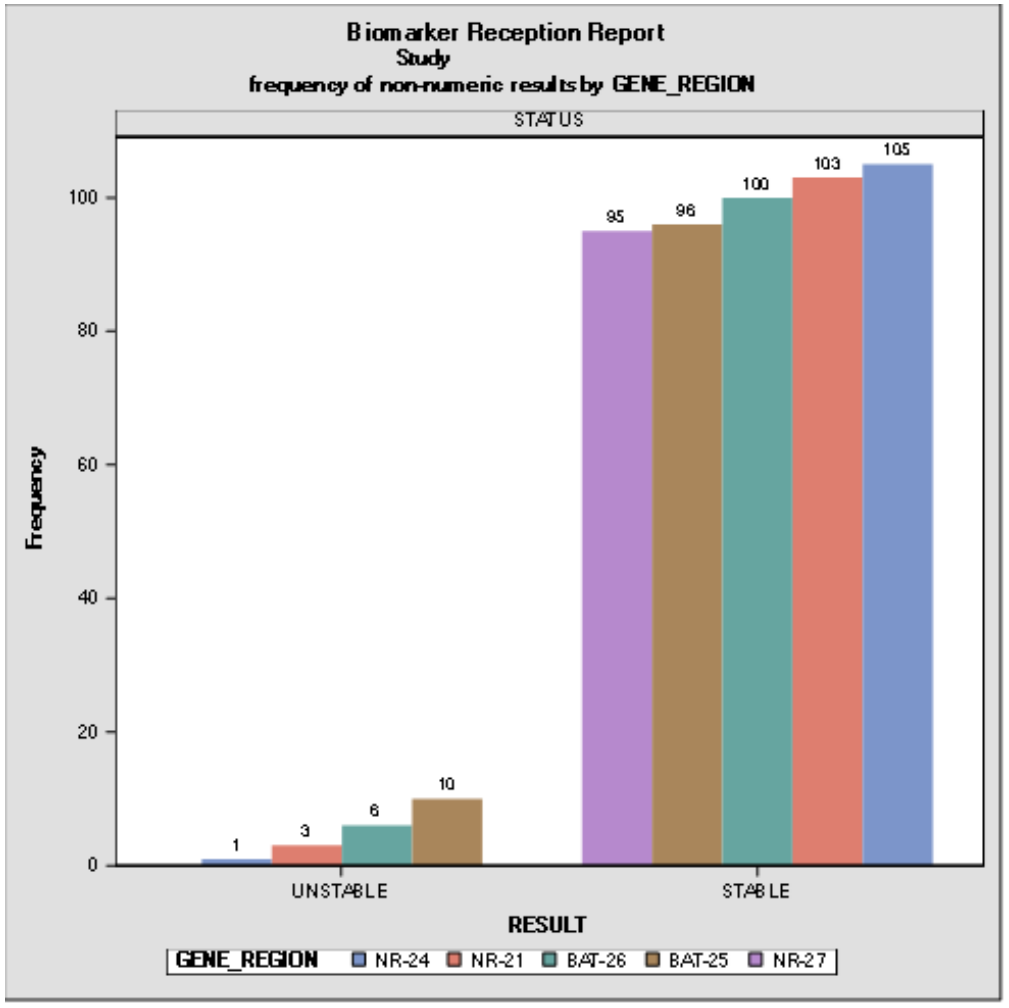
Item	Status
- Column GFMETHOD ('DROPLET DIGITAL PCR')	OK
- Column GFANMETH ('AVERAGE' 'CD274/GUSB RATIO')	OK
Checking the content of numerical columns:	
- Column GFORRES (min= 0.00, max= , otherval='')	OK
condition= GFTSTDTL='NORMALIZED LEVEL'	
(Descriptive statistics: n=578, MIN=22.2, MAX=213000, MEAN=9386.367128, MEDIAN=6840, STD=13674.132867)	
Other values: : 2 (0.3 %)	
- Column GFORRES (min= 0.00, max= , otherval='')	OK
condition= GFTSTDTL='SCORE'	
(Descriptive statistics: n=289, MIN=0.07277, MAX=37.05628, MEAN=1.5505060208, MEDIAN=0.4607, STD=3.8996037954)	
Other values: : 1 (0.3 %)	
- Column GFREPNUM (min= 1.00, max= , otherval='')	OK
(Descriptive statistics: n=870, MIN=1, MAX=1, MEAN=1, MEDIAN=1, STD=0)	
- Confirm the length of SUBJID values = 7	OK

Communication with Biomarker Team and Vendor

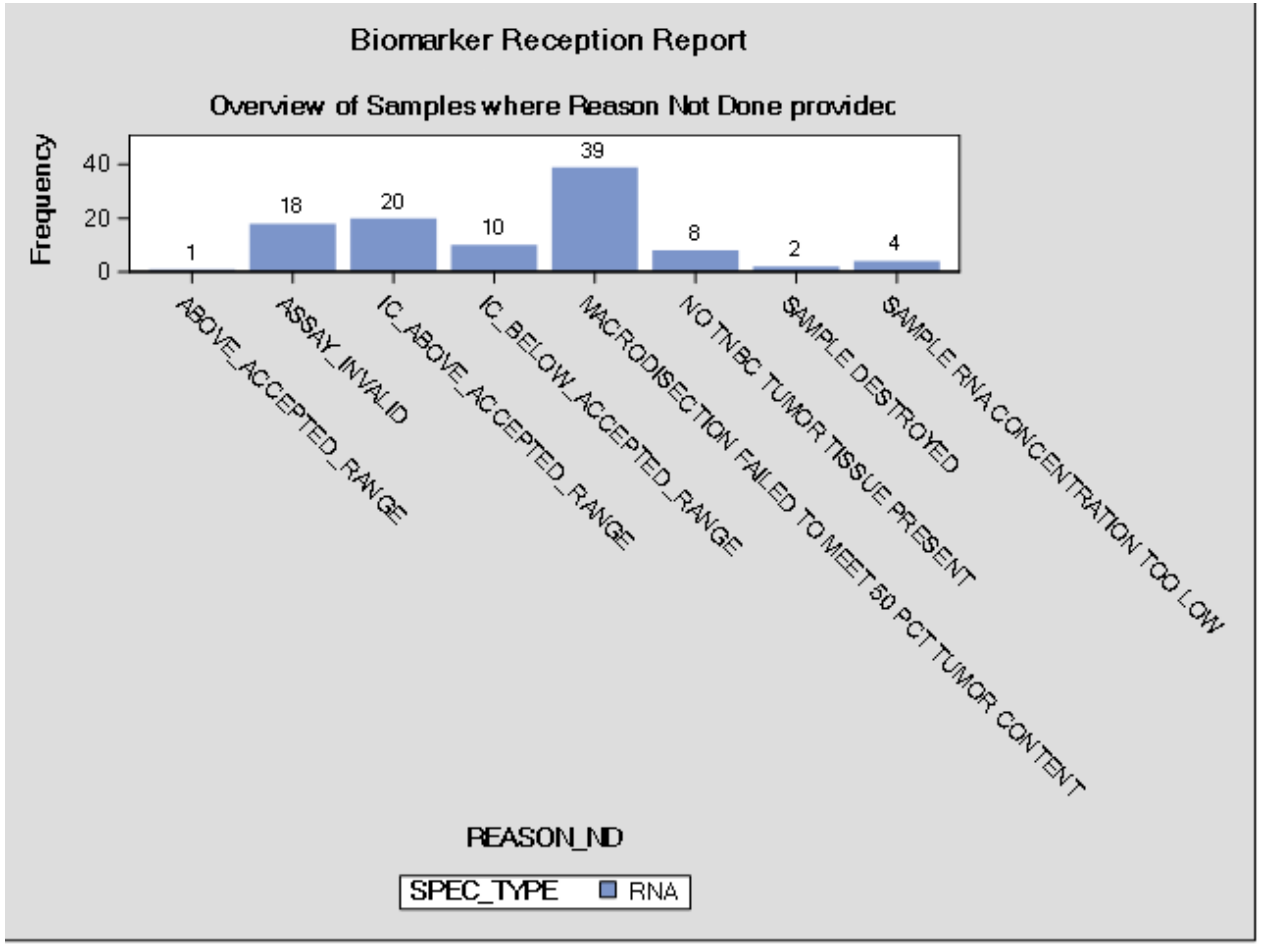
Data Quality Check Report

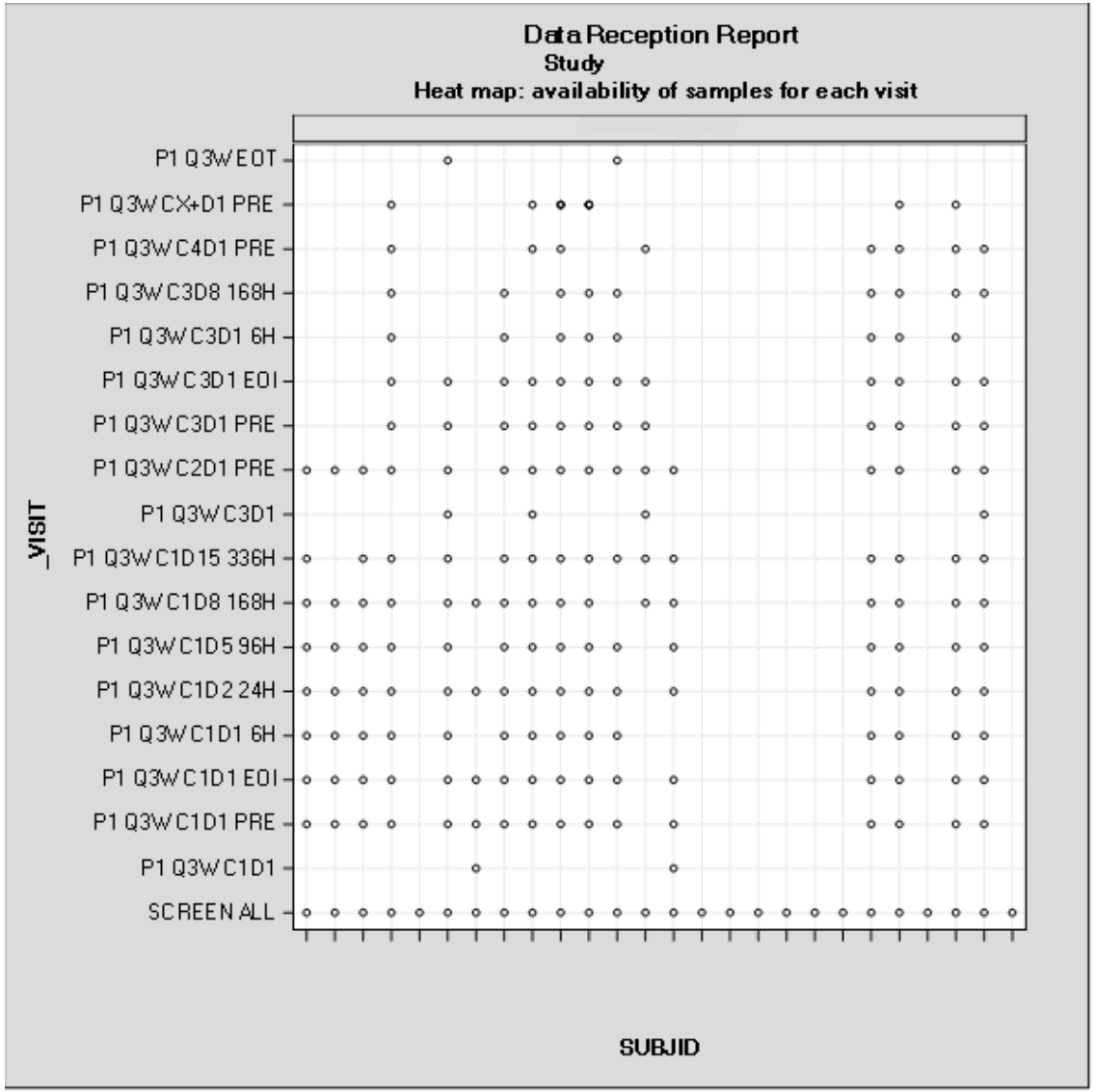
Data visualization





Data Quality Check Report





Data Issues Report

Collection of data issues and comments

	A	B	C	D
1	ITEM	STATUS	MERCK_KGAA_COMMENT	<VENDOR>_COMMENT
2	(Sample information file)	Shipping_manifest_1.csv, Shipping_manifest_2.csv		
3				
4	Column selected to compare:	COLSMPID		
5	Total of matching samples	290		
6				
7	Checking sample information file columns included in the data file:			
8				
9	-Difference found between column GFDTC in the data and column COLLECTION_DATE_TIME_ISO8601 in the sample information file	<gfrefid>, GFDTC= <yyyy-mm-ddThh:mm:ss>, COLLECTION_DATE_TIME_ISO8601=<yyyy-mm-ddThh:mm:ss>	Correction required	
10	-Difference found between column GFDTC in the data and column COLLECTION_DATE_TIME_ISO8601 in the sample information file	<gfrefid>, GFDTC= <yyyy-mm-ddThh:mm:ss>, COLLECTION_DATE_TIME_ISO8601=<yyyy-mm-ddThh:mm:ss>	Correction required	
11				
12	Data integrity:			
13				
14	3 record(s) per sample	290 (100.0%)	Sample <gfrefid> failed for testing. A record including the relevant information is enough.	
15				
16	Missing values:			
17				
18	- Column GFRUNID, percent of missing per GFREFID	<gfrefid>: 100.00% of records	OK, sample failed for testing.	
19	- Column GFMQLREV, percent of missing per GFREFID	<gfrefid>: 100.00% of records	OK, sample failed for testing.	
20				
21				

Key Factors

Experience summary

1. Developers are users
2. Collaboration
3. Communication
4. Standardized work environment
5. Team retain
6. Documentation

(Multiple functions)

+

1. Dataset unique key defined by user
2. User depending definition of parameters

-

Summary

Starting Point

- Different internal Data Management capabilities (Multiple software & methods)
- Non-standardized biomarker assay data
- Different vendors with basic data management capabilities
- Unreconciled sample data information

Achievement

- Support the Data Managers to check the biomarker data.
- Few people manage several studies simultaneously
- Support Biostatistics in the data comprehension, format and cleaning.
- Clear communication with the different stakeholders.
- Support the Scientist to review the data.
- Facilitate the data QC documentation
- Reconciliation against sample shipment manifest

Vision

2024

- **CDISC documentation** develops for multiple lab findings domains (CP, GF, IS, …)
- Vendors with **Data Management Teams**
- Many **vendors open** to the CDISC terminology

…

- Create **friendly interface** to facilitate the use from more people.
- **More standardized macros** based on the new data standards.
- **Validated** code repository
- Check the terminology considering the **Value Level Metadata** and not as single term.
- Check the **sample attributes** comparing with the clinical database in addition to the shipping manifest.
- **Issue management** tool orientation

Merck KGaA – BIOANA library contributors

Violeta Labarta Beceiro

Glenn Beeman

Samaneh Sabiani

Aruna Jayaprakash

Oliver Karch

Biomarker Team: Subject Matter Expert, Companion Diagnostic Expert, Biomarker Lead, Biomarker Operation Management, …
Standards Team

Colm Smyth

Wafaa Jebert

Q&A