

DH11 - Unlocking the Secrets of Health:

A Unified Approach to Biomarker and Clinical Data Integration

Pavan Kumar Tatikonda, Takeda

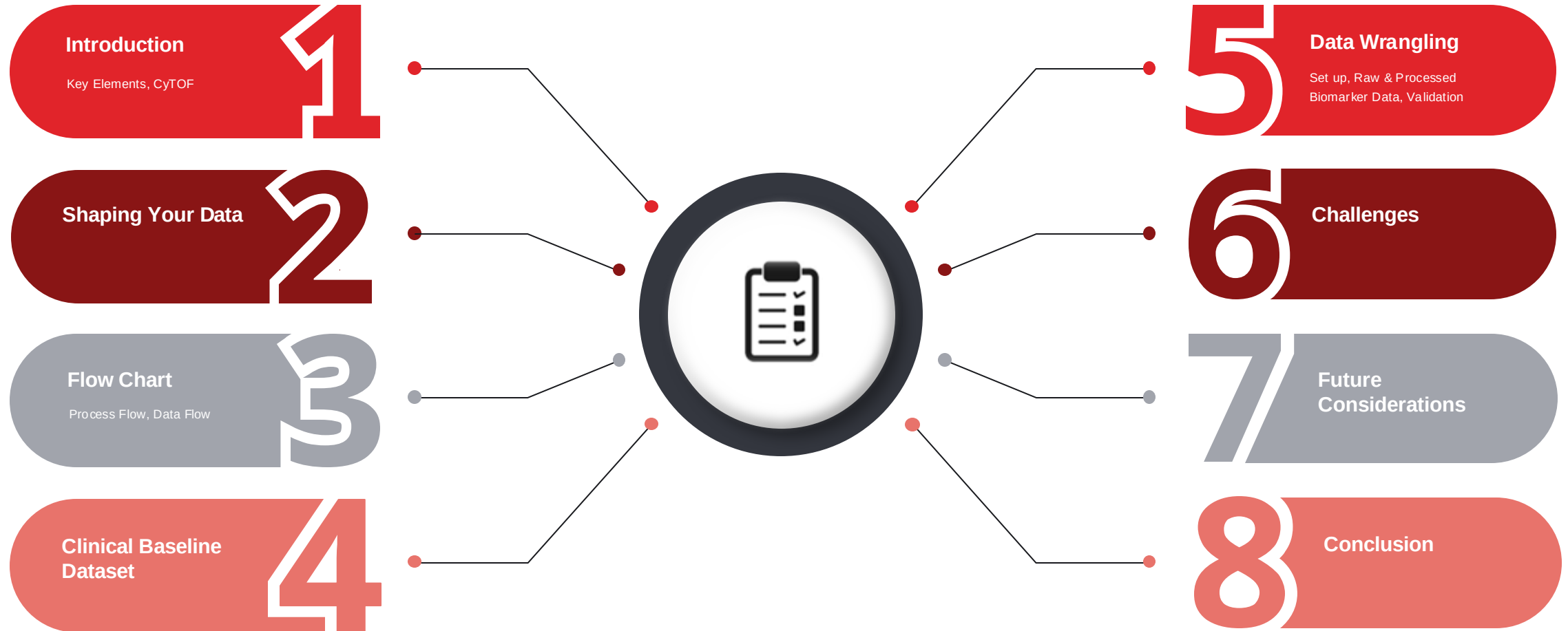
Xiaoyi Zhu, PPD

Sunithra Kadambi, Phastar

March 18, 2025



Agenda



➤ Overview of Biomarker-Driven Trials:

- Generate vast, high-dimensional data from multiple biospecimens and assay modalities.
- Widely adopted in clinical research:
 - ~50% of oncology trials.
 - 16.5% of all clinical trials incorporate biomarkers.

➤ Challenges:

- Linking data across modalities and centralizing reports for timely decision-making.
- Managing high-dimensional biomarker datasets.
- Manual processes (e.g., data requests via email, data stored and shared on SharePoint) and reliance on non-standard formats.
- Limited or non-existent validation methods.
- Reproducibility of R analytical code, workflows, and results.

➤ Call to Action:

- A Centralized Biomarker Database: Integrates clinical baseline data, biomarker data, and sample data from trial start.

➤ **Exploratory Analysis Plan (EAP):**

- Details key biomarker questions, hypotheses, and intended functions.
- Describes the analysis approach of the exploratory data.
- Aligns with exploratory objectives and endpoints outlined in the clinical trial protocol.

➤ **Clinical Baseline Dataset:**

- Integrates key demographics, baseline characteristics, safety, and efficacy variables.

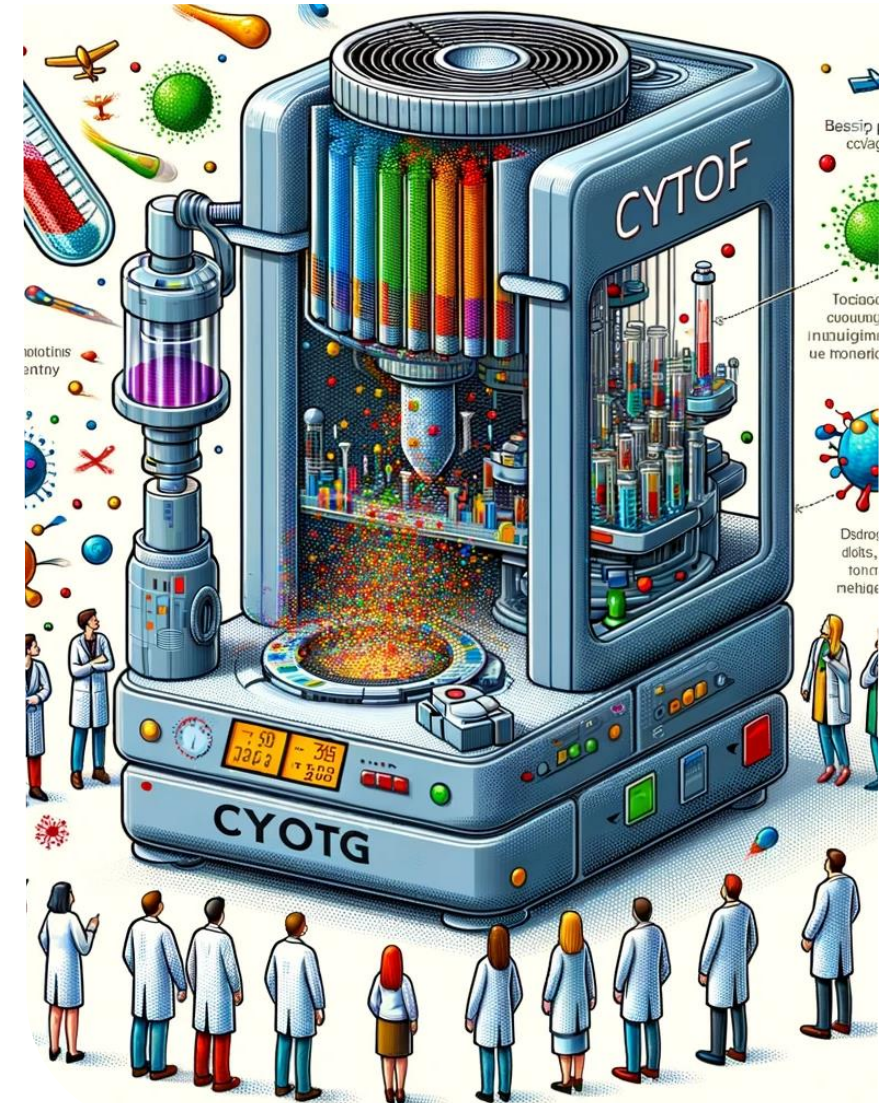
➤ **Data Processing and Storage:**

- Statistical software R handles large data volumes.
- Amazon S3 ensures secure storage and sharing of raw and processed data.
- R packages enable seamless data transfer to/from Amazon S3 bucket.

Key Elements cont.,

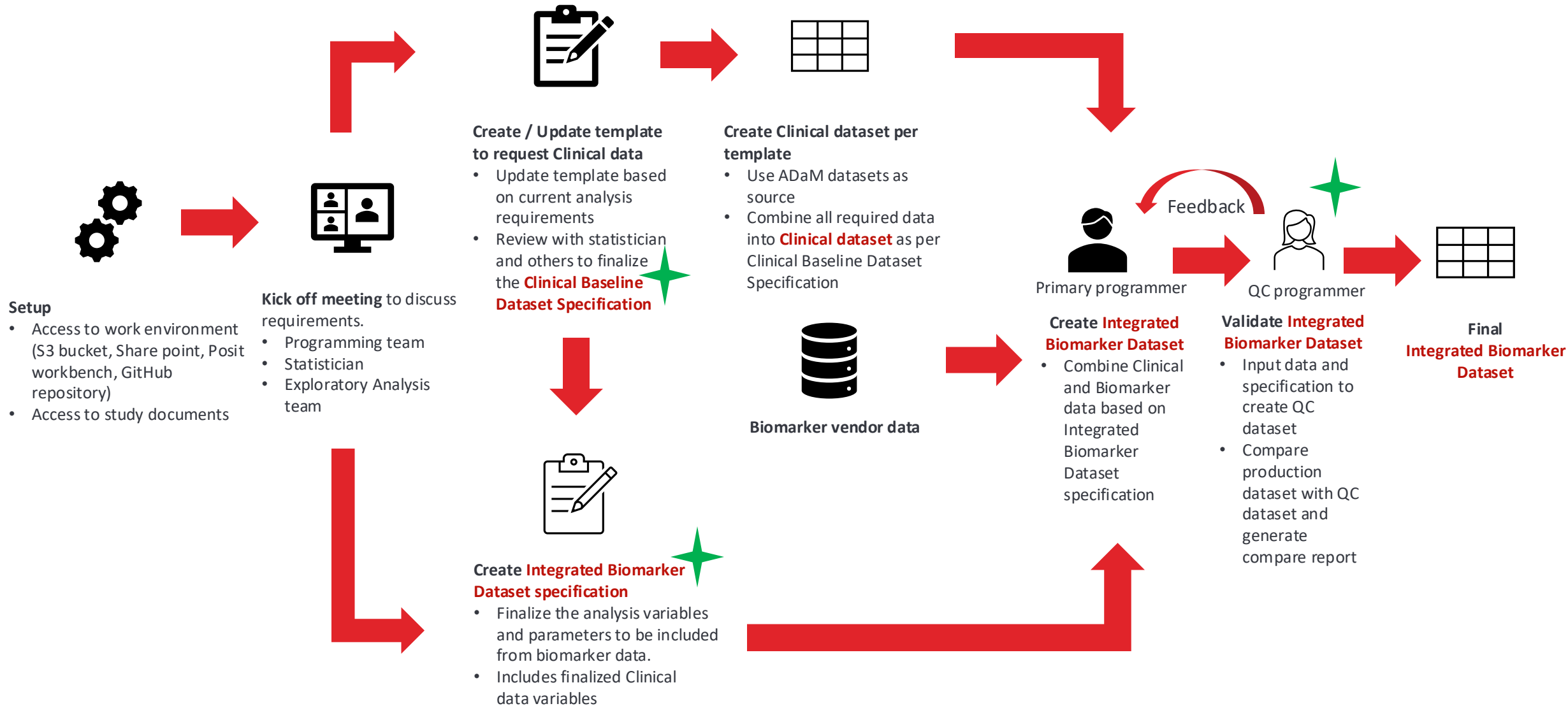
➤ High-priority Question:

- Does the baseline immune status (immunophenotype, cytokines; > 500 reportable) determine the ability of a patient to mount an anti-tumor immune response?
 - (e.g., do deficiencies in the patients' immune cells (frequency/activation/exhaustion) correlate with non-response?)
- Cytometry by time-of-flight (CyTOF) is a recently introduced technology that uses mass spectroscopy to measure the abundance of metal isotope labels on antibodies and other tags (such as peptide-MHC tetramers for labeling specific T cells) on single cells.
- The current dataset was produced by analyzing specimens such as RBC-lysed whole Blood and Bone Marrow Assay (BMA) using gate definitions and markers, produced results such as Mean, Percent of parent Populations results along with event counts, Leukocytes Counts, etc.,
- Our work used CyTOF data; we can extend it to all other Biomarker data.

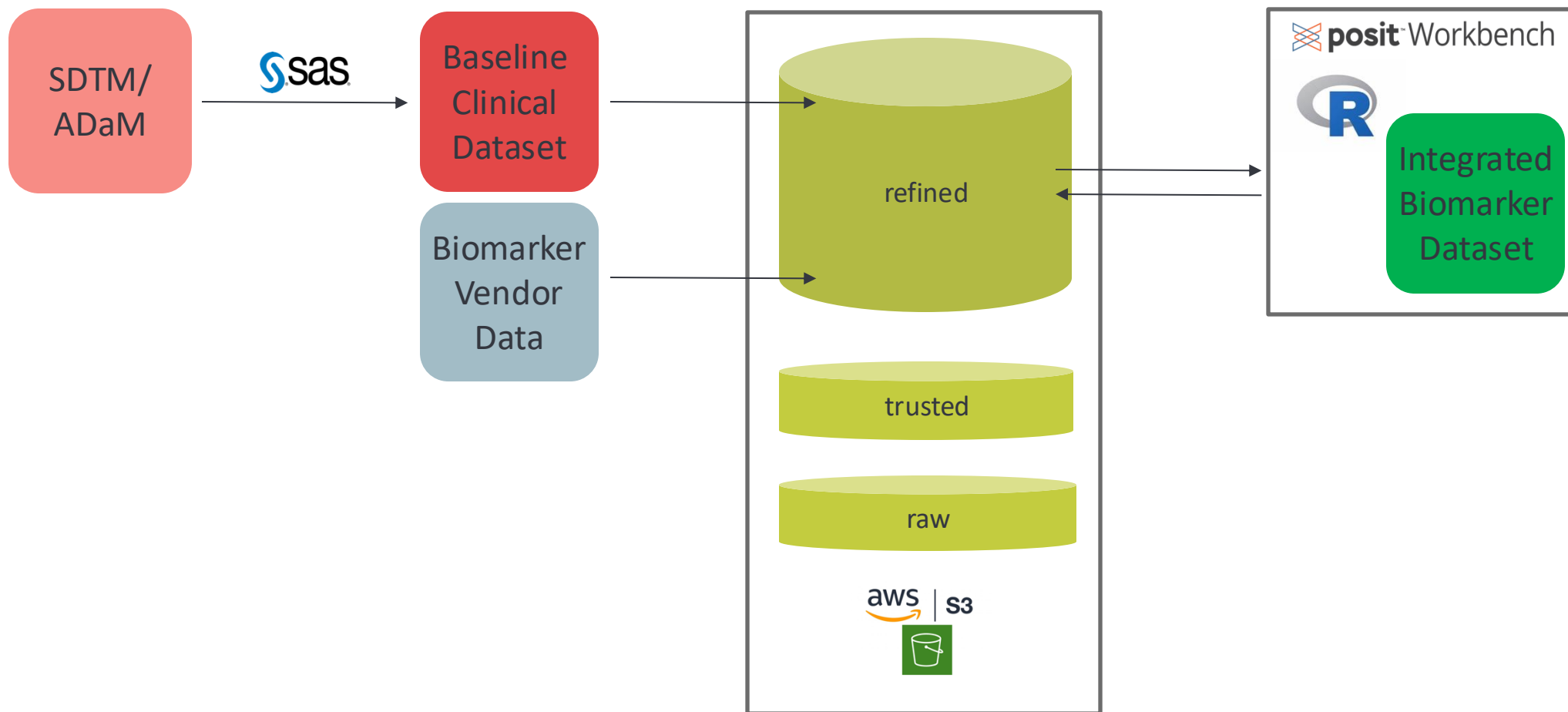


- What if you were asked to produce an exploratory analysis dataset that involves high-dimensional biomarker data?
- Would you opt for a long or wide format?
- In a wide format, results for various biomarkers across visits are presented as columns, making them visually interpretable.
- In a long format, it organizes data over multiple rows, enabling easier aggregated analysis.
- Biomarker characteristics, gate definitions, marker strings, source samples, and reagents combine to define analysis parameters like 'Mean Metal Intensity,' 'Event Counts,' and 'Percent of Leukocytes'.

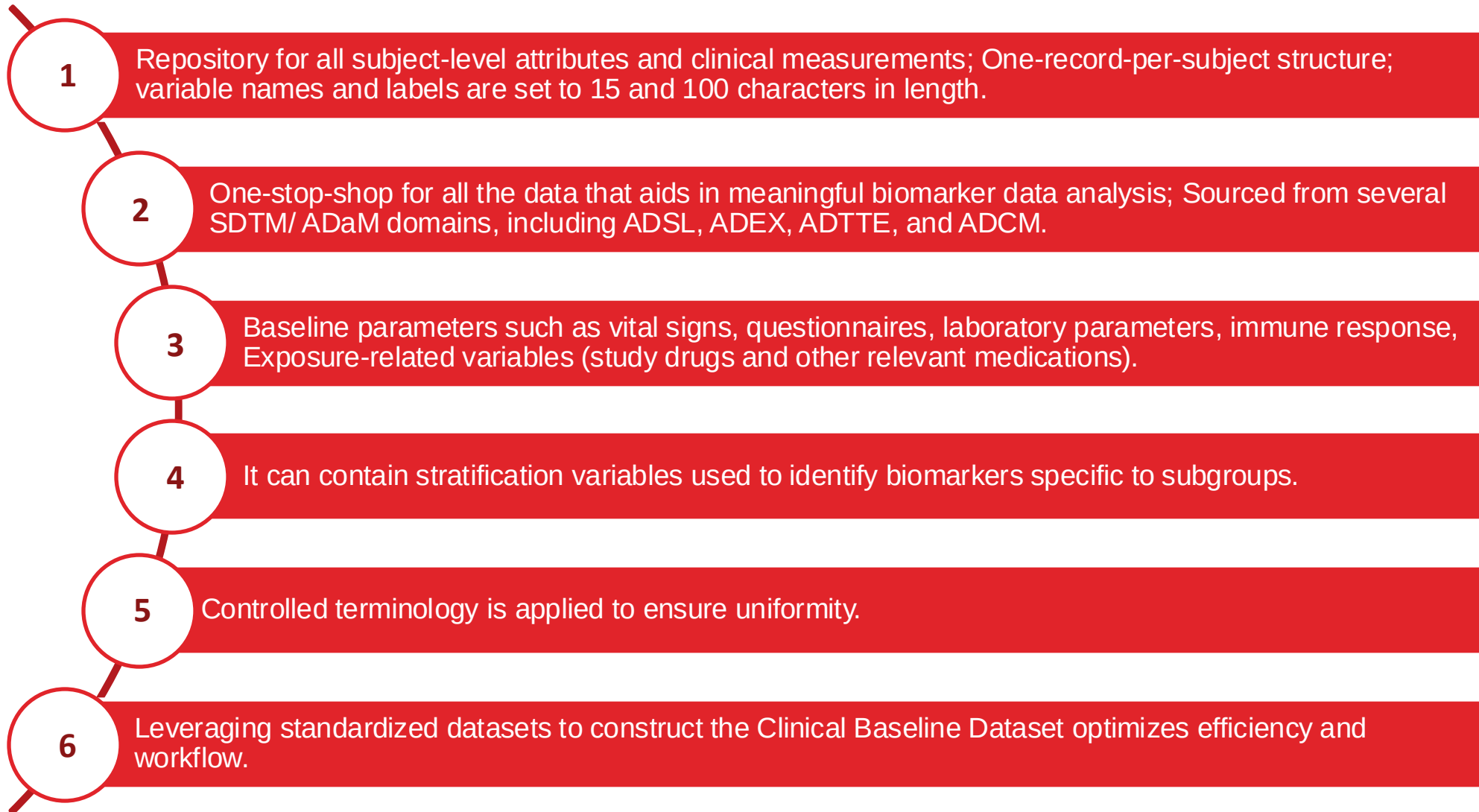
Process Flow



Data Flow



Clinical Baseline Dataset



Clinical Baseline Dataset Cont.,



Dataset	Order	Domain	Variable Name	Variable Label	Type	Expected	Takeda Core	Controlled Terms	Format	Origin	Predecessor Name	Derivation
CLINDSBL	1	Demo	USUBJID	Unique Subject Identifier	Char	50	Req			Predecessor	ADSL.USUBJID	
CLINDSBL	2	Demo	TRT01P	Planned Treatment for Period 01	Char	100	Req			Predecessor	ADSL.TRT01P	
CLINDSBL	3	Demo	TRT01A	Actual Treatment for Period 01	Char	100	Req			Predecessor	ADSL.TRT01A	
CLINDSBL	4	Demo	AGE	Age	Num	8	Req			Predecessor	ADSL.AGE	
CLINDSBL	5	Demo	AGEGRP	Age Group	Char	50	Perm	(< 65 years, >= 65 years)		Derived		
CLINDSBL	6	Demo	SEX	Sex	Char	10	Req	(Female, Male)		Predecessor	ADSL.SEX	
CLINDSBL	7	Demo	RACE	Race	Char	50	Req	AMERICAN, WHITE, NOT REPORTED)		Predecessor	ADSL.RACE	

Dataset	Order	Domain	Variable Name	Variable Label	Type	Expected	Takeda Core	Controlled Terms	Format	Origin	Predecessor Name	Derivation
CLINDSBL	8		ECOGCAT	ECOG Category	Char	10	Perm	(0-1, 2, Missing)		Derived		
CLINDSBL	9		CRCL	Calculated Creatinine Clearance (mL/min)	Num	8	Perm			Derived		
CLINDSBL	10		CRCLCAT	Calculated CrCl (mL/min) Category	Char	50	Perm	(< 60 mL/min, >= 60 mL/min)		Derived		
CLINDSBL	11		IGATYPE	IgA Immunoglobulin Type	Char	1	Perm	(NY)		Derived		
CLINDSBL	12		HCHTYP	Heavy Chain Type	Char	10	Perm	(IgG, IgA, IgD, IgE, IgM...)		Derived		
CLINDSBL	13		LCHTYP	Light Chain Type	Char	10	Perm	(Kappa, Lambda, Unknown)		Derived		
CLINDSBL	14		CD38E90D	Exp to Anti-CD38 w/i 90 Days Prior Trt	Char	1	Perm	(NY)		Derived		
CLINDSBL	15		IMIDE60D	Exp to IMiD Within 60 Days Prior Trt	Char	1	Perm	(NY)		Derived		
CLINDSBL	16		PIE60D	Exposed to Proteasome Inhibitor Within 60 D	Char	1	Perm	(NY)		Derived		
CLINDSBL	17		ADABL	ADA Status at Baseline	Char	10	Perm	(Positive, Negative, Missing)		Derived		
CLINDSBL	18		CYTOPFR	Cytogenetics Testing Performed	Char	1	Perm	(NY)		Derived		
CLINDSBL	19		CYTOHR	High Risk Cytogenetics	Char	10	Perm	(N,Y, Not Done)		Derived		
CLINDSBL	20		CADEL13	Chromosomal Aberration: del 13	Char	200	Perm			Derived		
CLINDSBL	28		CA1Q21	Chromosomal Aberration: 1q21	Char	200	Perm			Derived		
CLINDSBL	29		CAHYPER	Chromosomal Aberration: hyperdiploidy	Char	200	Perm			Derived		
CLINDSBL	30		CAHYPER	Chromosomal Aberration: hyperdiploidy	Char	200	Perm			Derived		

Dataset	Order	Domain	Variable Name	Variable Label	Type	Expected	Takeda Core	Controlled Terms	Format	Origin	Predecessor Name	Derivation
CLINDSBL	36		RESPFLIRF	Responder Flag per IRF	Char	1	Perm	(NY)		Derived		
CLINDSBL	37		DORDINV	Dur of Response per INV (days)	Num	8	Req			Derived		
CLINDSBL	38		DORDIRF	Dur of Response per IRF (days)	Num	8	Req			Derived		
CLINDSBL	39		DORMINV	Dur of Response per INV (months)	Num	8	Req			Derived		
CLINDSBL	40		DORMIRF	Dur of Response per IRF (months)	Num	8	Req			Derived		
CLINDSBL	41		DORCNSINV	Dur of Response Censor per INV	Num	8	Req			Derived		
CLINDSBL	42		DORCNSIRF	Dur of Response Censor per IRF	Num	8	Req			Derived		
CLINDSBL	43		PFS DINV	PFS per INV (days)	Num	8	Req			Derived		
CLINDSBL	44		PFS DIRF	PFS per IRF (days)	Num	8	Req			Derived		
CLINDSBL	45		PFS MINV	PFS per INV (months)	Num	8	Req			Derived		
CLINDSBL	46		PFS MIRF	PFS per IRF (months)	Num	8	Req			Derived		
CLINDSBL	47		PFS CNSINV	PFS Censor per INV	Num	8	Req			Derived		
CLINDSBL	48		PFS CNSIRF	PFS Censor per IRF	Num	8	Req			Derived		

The Data Wrangling Journey



- Setup working environment for programming.
 - Created a working repository on GitHub for the biometrics team to store and share files.

File Structure:

- `install_packages.R` : installs all required R packages
- `start.R` : provides starter code for connecting Posit/RStudio with S3 bucket
- `compare.R` : compares our results with the original output files on S3 bucket
- `CytoF_Data_Process_MMI_Blood_BMA_Long.R` : cleans and combines the biomarker MMI data with baseline clinical dataset
- `CytoF_Data_Process_Percent_Blood_BMA_Long.R` : cleans and combines the biomarker PERCENT data with baseline clinical dataset
- `\qc` folder: stores QC programs and comparison report
- `\archive` folder: stores archived files for versioning

The Data Wrangling Journey Cont.,



- Link Posit Workbench to the GitHub repository for collaboration and version control purposes.

Here are step-by-step instructions for setting up an SSH key within the Posit Workbench terminal:

1. Open the Posit Workbench terminal: Launch Posit Workbench and navigate to the terminal interface. You can usually find it in the toolbar at the top of the screen, or you can access it through the "Tools" menu.
2. Check for existing SSH keys: To determine if you already have SSH keys set up, enter the following command in the terminal:

```
1 ls ~/.ssh
```

This will list any existing SSH keys in the `~/.ssh` directory.



9. The last step in this section is copying the public key: To use the SSH key, you need to copy the public key to the remote Github server. Enter the following command to display the contents of your public key:

```
1 cat ~/.ssh/id_ed25519.pub
```

The output will be the public key text. You can copy this for the next section.

That's it! You have now set up an SSH key within the Posit Workbench terminal and can use it to securely connect to remote servers.

- Get access to the Amazon S3 bucket.
- Link Amazon S3 bucket with Posit Workbench.
- Process-related README files are produced for future use in standardization.

Biomarker Raw Data



- Vendor-produced raw datasets:
 - By analyzing specimens of RBC-lysed whole Blood and Bone Marrow Assay (BMA).
 - Results, such as mean metal intensity, Percentage of parent populations, event counts, and leukocyte counts, were produced using gate definitions and markers.
- The raw biomarker dataset consists of Subject ID (SUBJID), Gate Definition (LBGATDEF), Marker String (LBMRKSTR), Test Detail (LBTSTDTL), Specimen, and reagent-related variables.
- The dataset's structure is one record per subject per biomarker, test, and timepoint.

LBSBMRKS	LBGATDE	LBMRKSTR	LBTSTDTL	REAGMMI	LBORRES	LBSPEC	LBMETHOD	PCTLEUK	EVNTCT	VISIT
mDC	mDC	CD14-CD16-CD56-CD3-CD19-	mean	162Dy_CD86	16.46166151	RBC Lysed Whole Blood	CyTOF	0.116477285	274	V21 C5D1
mDC	mDC	CD14-CD16-CD56-CD3-CD19-	mean	172Yb_CD83	0.435098903	RBC Lysed Whole Blood	CyTOF	0.116477285	274	V21 C5D1
pDC	pDC	CD14-CD16-CD56-CD3-CD19-	mean	162Dy_CD86	5.920483597	RBC Lysed Whole Blood	CyTOF	0.017003983	40	V21 C5D1
pDC	pDC	CD14-CD16-CD56-CD3-CD19-	mean	172Yb_CD83	0.28795932	RBC Lysed Whole Blood	CyTOF	0.017003983	40	V21 C5D1
mDC	mDC	CD14-CD16-CD56-CD3-CD19-	mean	162Dy_CD86	16.47284876	RBC Lysed Whole Blood	CyTOF	0.138468238	283	V2 Baseline
mDC	mDC	CD14-CD16-CD56-CD3-CD19-	mean	172Yb_CD83	0.83951769	RBC Lysed Whole Blood	CyTOF	0.138468238	283	V2 Baseline
pDC	pDC	CD14-CD16-CD56-CD3-CD19-	mean	162Dy_CD86	6.12556888	RBC Lysed Whole Blood	CyTOF	0.036696529	75	V2 Baseline
pDC	pDC	CD14-CD16-CD56-CD3-CD19-	mean	172Yb_CD83	0.810534398	RBC Lysed Whole Blood	CyTOF	0.036696529	75	V2 Baseline

Variable Name	Description
SUBJID	Subject ID
LBREFID	Sample ID
LBGATDEF	Gate Definition
LBMRKSTR	Marker String
LBTSTDTL	Test Detail (e.g., "mean" or "percent")
REAGMMI	Metallic Based Mean Intensity Reagent
LBSPEC	Specimen Type
VISIT	Visit
LBORRES	Original Result
PCTLEUK	Percent of Leukocytes
EVNTCT	Event Count

- **Integrated biomarker dataset specification** was developed to guide the creation of the final dataset.
- **ADaM BDS variables like** PARAM, PARAMCD, BASE, CHG, and other time points and biomarker-identifying variables are derived. A few variables were remapped, E.g., visit, to apply study-specific controlled terminology.
- Some variables retained the same names as the source data, such as LBGATDEF, LBMRKSTR, and REAGMMI, to ensure end-user interpretability.
- TESTDETAIL was created to aid efficient data screening by concatenating variables LBGATDEF, LBMRKSTR, and LBTSTDTL.
- **‘Event count’ and ‘Percentage of leukocytes’ were additional results maintained in wide format.** The result variables were named EVNTCTORIGRES and PCTLEUKORIGRES.
- The structured dataset design optimizes **compatibility with statistical and visualization tools**, allowing for more effective cross-functional analyses in a regulated clinical research environment.

Processed Biomarker Data Cont.,



LBGATDEF	LBMRKSTR	LBTSTDTL	TESTDETAIL	REAGMMI	PARAMCD	PARAM	ORIGRES	BASE	CHG	FCHG	VNTCTO	IGRES	EVNTCTBASE	EVNTCTCHG	EVNTCTFCHG	PCTLEUKORIG	RES	PCTLEUKBASE	PCTLEUKCHG	PCTLEUKFCHG
mDC	CD14-CD16	mean	mDC CD14-CD16-CD56-CD3-CD19- mean	162Dy_CD86	MMIMDC86	Mean Metal Int	16.46166	16.47285	-0.01119	0.999320868	274	283	-9	0.9681978	0.116477285	0.138468238	-0.021990953	0.841184134		
pDC	CD14-CD16	mean	pDC CD14-CD16-CD56-CD3-CD19- mean	172Yb_CD83	MMIPDC83	Mean Metal Int	0.287959	0.810534	-0.52258	0.355270943	40	75	-35	0.533333333	0.017003983	0.036696529	-0.019692546	0.463367611		
Classical monocy	HLADR+	mean	Classical_monocytes HLADR+ mean	162Dy_CD86	MMICM86	Mean Metal Int	21.10263	16.1503	4.952338	1.3066407	171699	130915	40784	1.31153038	72.98917271	64.05501544	8.934157277	1.139476311		
Non-classical monocy	HLADR+	mean	Non-classical_monocytes HLADR+ mean	172Yb_CD83	MMINCM83	Mean Metal Int	0.296046	0.674158	-0.37811	0.439134715	66	257	191	0.25690922	0.028056572	0.125746774	-0.097690202	0.223119618		

Clinical
Baseline
Variables



LBGATDEF	LBMRKSTR	LBTSTDTL	TESTDETAIL	REAGMMI
mDC	CD14-CD16-CD56-CD3-CD19-	mean	mDC CD14-CD16-CD56-CD3-CD19- mean	162Dy_CD86
mDC	CD14-CD16-CD56-CD3-CD19-	mean	mDC CD14-CD16-CD56-CD3-CD19- mean	172Yb_CD83
pDC	CD14-CD16-CD56-CD3-CD19-	mean	pDC CD14-CD16-CD56-CD3-CD19- mean	162Dy_CD86
pDC	CD14-CD16-CD56-CD3-CD19-	mean	pDC CD14-CD16-CD56-CD3-CD19- mean	172Yb_CD83

TESTDETAIL	REAGMMI	PARAMCD	PARAM	ORIGRES
mDC CD14-CD16-CD56-CD3-CD19- mean	162Dy_CD86	MMIMDC86	Mean Metal Intensity of mDC on CD86	16.46166
pDC CD14-CD16-CD56-CD3-CD19- mean	172Yb_CD83	MMIPDC83	Mean Metal Intensity of pDC on CD83	0.287959
Classical_monocytes HLADR+ mean	162Dy_CD86	MMICM86	Mean Metal Intensity of Classical Monocytes on CD86	21.10263
Non-classical_monocytes HLADR+ mean	172Yb_CD83	MMINCM83	Mean Metal Intensity of Non-classical Monocytes on CD83	0.296046

ORIGRES	BASE	CHG	FCHG	VNTCTO	IGRES	EVNTCTBASE	EVNTCTCHG	EVNTCTFCHG	PCTLEUKORIG	RES	PCTLEUKBASE	PCTLEUKCHG	PCTLEUKFCHG
16.46166	16.47285	-0.01119	0.999320868	274	283	-9	0.9681978	0.116477285	0.138468238	-0.021990953	0.841184134		
0.287959	0.810534	-0.52258	0.355270943	40	75	-35	0.533333333	0.017003983	0.036696529	-0.019692546	0.463367611		
21.10263	16.1503	4.952338	1.3066407	171699	130915	40784	1.31153038	72.98917271	64.05501544	8.934157277	1.139476311		
0.296046	0.674158	-0.37811	0.439134715	66	257	191	0.25690922	0.028056572	0.125746774	-0.097690202	0.223119618		

Important processed variables

- To keep transparency, carry forward the biomarker identification variables.
- Derived Timepoint, PARAMCD, PARAM, TESTDETAIL variables.
- The primary results are the ORIGRES variable and derived variables, such as BASE, CHG, and FCHG, which increase efficiency across time points and test types.
- Derived for all corresponding result variables, such as Event Count and Percent of Leukocytes.

Variable Name	Description
SUBJID	Subject ID
SAMPLEID	Sample ID
LBGATDEF	Gate Definition
LBMRKSTR	Marker String
LBTSTDTL	Test Detail (e.g., “mean” or “percent”)
REAGMMI	Metallic Based Mean Intensity Reagent
TESTDETAIL	Combined Value of Gate Definition Marker String Test Detail
SPECIMENTYP	Specimen Type
TIMEPOINT	Timepoint
PARAM	Parameter
PARAMCD	Parameter Code
ORIGRES	Original Result
BASE	Baseline Result
CHG	Change from Baseline
FCHG	Fold Change over Baseline
EVNTCTORIGRES	Event Count Original Result
EVNTCTBASE	Event Count Baseline Result
EVNTCTCHG	Event Count Change from Baseline
EVNTCTFCHG	Event Count Fold Change over Baseline
PCTLEUKORIGRES	Percent of Leukocytes Original Result
PCTLEUKBASE	Percent of Leukocytes Baseline Result
PCTLEUKCHG	Percent of Leukocytes Change from Baseline
PCTLEUKFCHG	Percent of Leukocytes Fold Change over Baseline

Data Validation



- A parallel **programmer validates** the Integrated Biomarker Dataset.
- Independently **recreate the dataset** using source datasets and specifications.
- Performed in a validated R environment.
- Checks **completeness, consistency, and adherence to specification** of both data and metadata.
- Resolves discrepancies between primary and QC datasets.
- Clean **compare report** marks a finalized Integrated Biomarker Dataset.

Understanding Biomarker Data

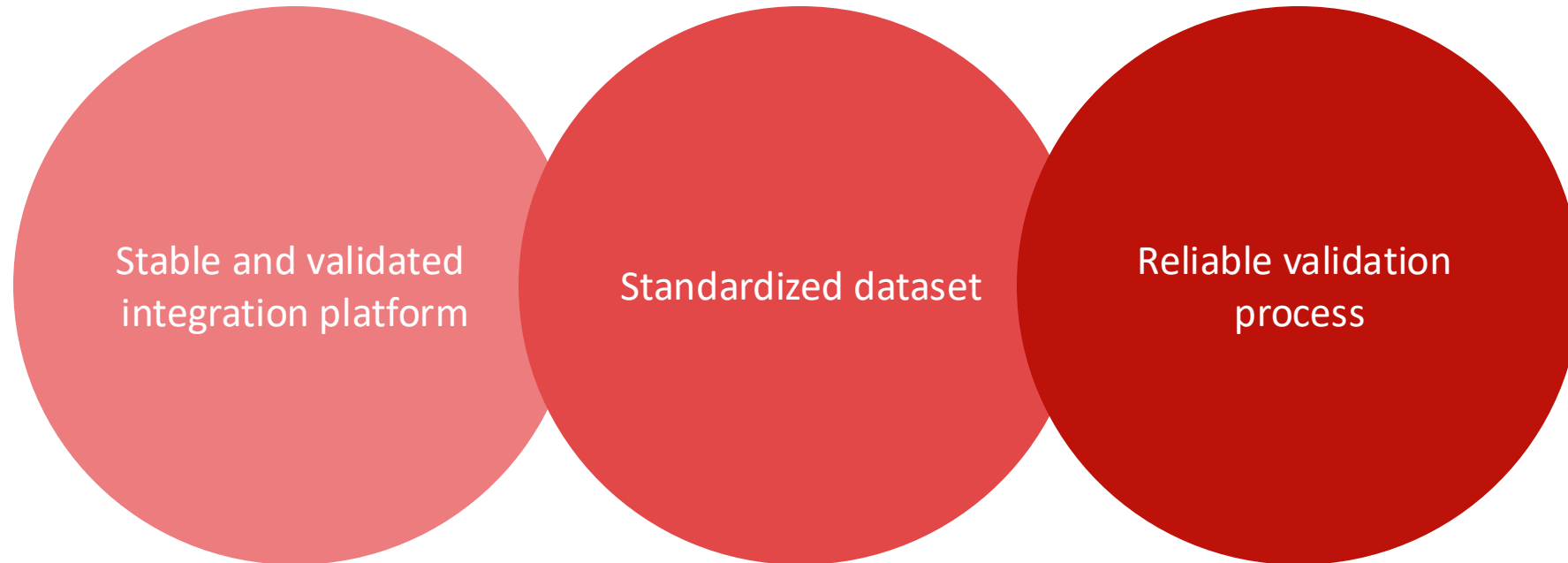
- Biometrics team has limited knowledge about exploratory biomarker data.
- Multiple consultations with Biomarker data SMEs.
- Knowledge sharing early on to prevent downstream delays.

Decision on Data Orientation

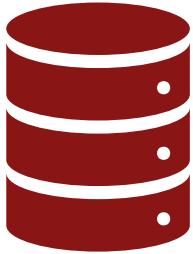
- The exploratory team is used to working with wide-format data, and standardizing in this orientation is challenging.
- Variability in the number of biomarkers.
- Create a long format dataset (CDISC ADaM-like) for Exploratory analysis.

- Depending on the data volume that needs processing, the programming team can **consider using SAS or other programming languages instead of R.**
- The **R Shiny app can partly automate** the process of requesting data and creating analysis datasets.
- Developing an R shiny app for **data visualization and interactive reports** can be explored for established biomarkers.

Our work thus **establishes some critical elements** of creating a biomarker dataset for exploratory analysis.

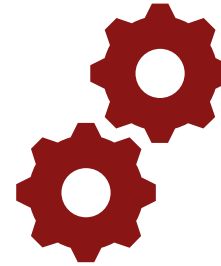


Establishing a centralized biomarker database using a streamlined process is paramount for effective data analysis and exploratory research.



Integrated Biomarker Dataset

- Robust data management
- Enhanced decision making
- Operational efficiency



Streamlined Integration Process

- Improved quality control (QC)
- Data integrity and reliability
- Enhanced collaboration

Acknowledgements



We thank our cross-functional colleagues for their guidance and support during the conception and implementation of the biomarker data integration process. We also thank Maria Dalton, Venky Chakravarthy, Samanthi Perera, and Kevin Galinsky for their valuable feedback on this paper.

Conflicts of Interest



The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Contact Information



Author Name: Pavan Kumar Tatikonda

Company: Takeda Development Center Americas Inc.,

Email: pavan.tatikonda@takeda.com

Author Name: Xiaoyi Zhu

Company: PPD, part of Thermo Fisher Scientific

Email: sherry.zhu1@thermofisher.com

Author Name: Sunithra Kadambi

Company: Phastar Inc.

Email: sunithra.kadambi@phastar.com

THANK YOU !!

Q & A