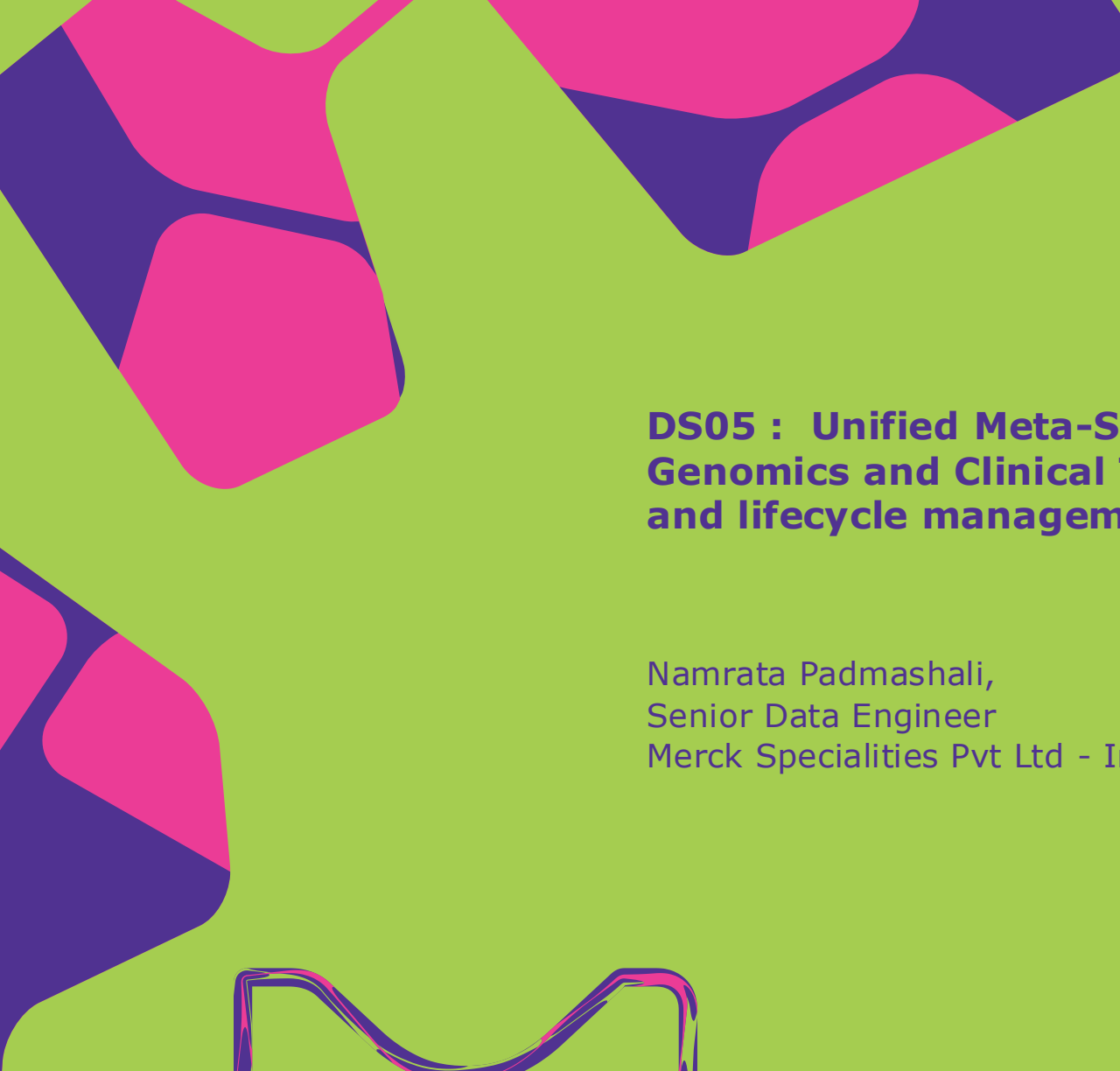
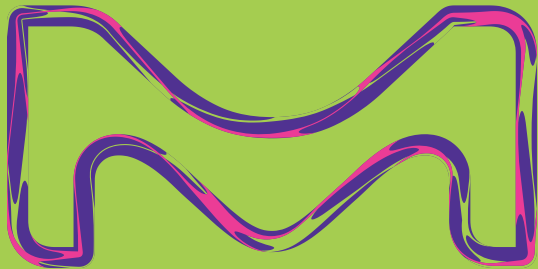




DS05 : Unified Meta-Study: Harmonizing Clinical Biomarker, Genomics and Clinical Trial Data for enhanced access, usability, and lifecycle management

Namrata Padmashali,
Senior Data Engineer
Merck Specialities Pvt Ltd - India



MERCK

Agenda

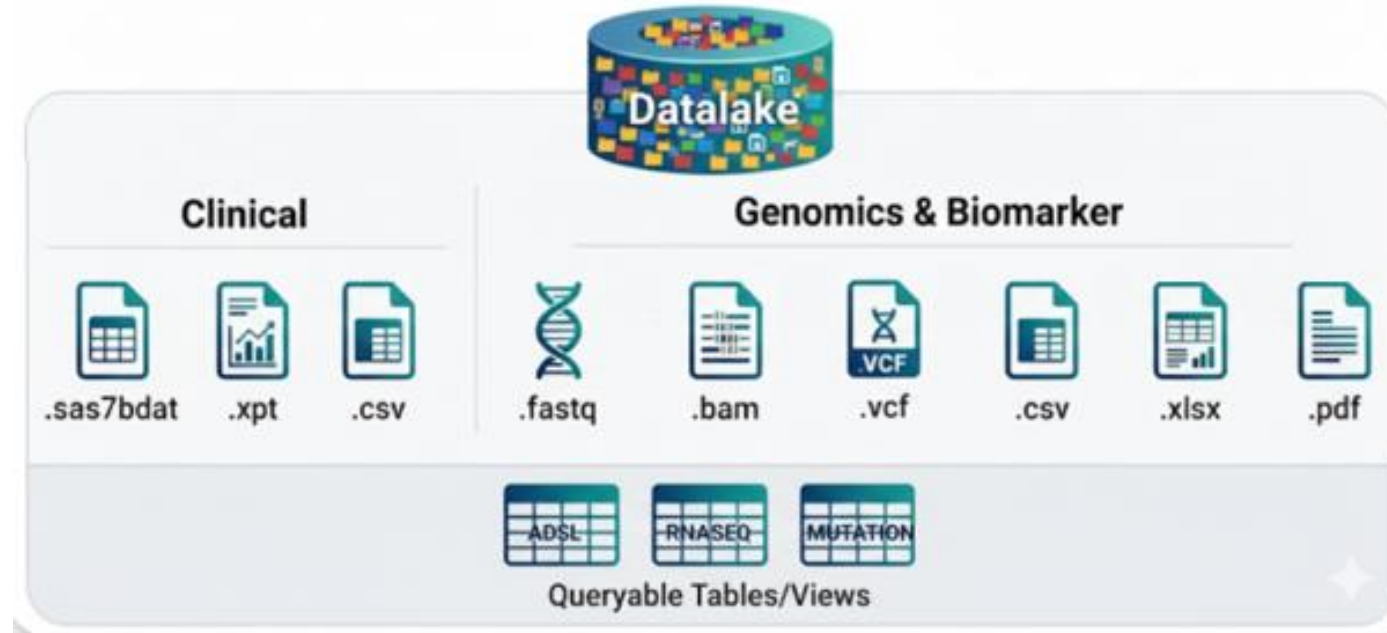
-
- 01** The Data Opportunity & Landscape
 -
 - 02** Data Hunt Problem & Data Product as a Solution
 -
 - 03** Harmonization Challenges & Strategy
 -
 - 04** Harmonized Master Schema
 -
 - 05** Use Case + Key Takeaways
 -

The Data Opportunity

Scientist A Target Discovery



P2026 - 12 clinical trials across 4 indications



Scientist B Oncology R&D



We need data from all P2026 trials to support hypothesis generation for *compound AB52* - integrated and harmonized data including survival rates, tumor stage, genomic deletions (chr2:2p25) and few more biomarkers

We need specific clinical and biomarker parameters from selected P2026 trials for drug repositioning – demographics, response rates, PFS, KRAS & EGFR mutations, PD-L1 expression, tumor type and disease type

Problem : The Data Hunt

Fragmented Data

- 12 folders per trial
- Visit each study folder to gather data
- Verify if files and views match latest cutoff dates.
- Decode complex variable names

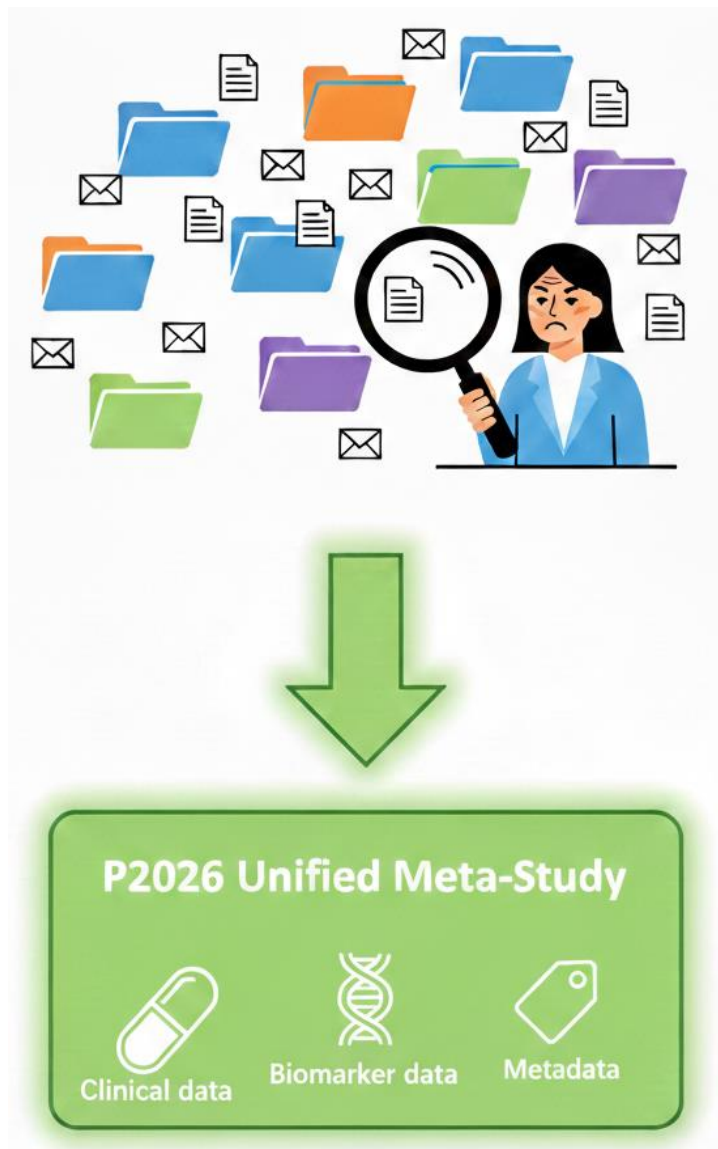
Messy Variables

- Different variable name
(e.g. *BLECOG* vs *ECOGBL*,
pos_chrom vs *pos.chrom*)
- Different biomarker
output layouts (varies by
assay, panel, vendor, etc.)
- Data trapped in PDFs

Time Sink

- Siloed & repeated
harmonization efforts
- Clinical IDBs might miss
key analysis variables
- No harmonized genomics
outputs per compound
- Limited resources

Solution : Unified Meta-Study Data Product



- ✓ **Unified Meta-Study:** All P2026 studies combined
- ✓ **Data Types:** Clinical + Biomarker + Metadata
- ✓ **Solves fragmentation:** Harmonized & reusable
- ✓ **Empowers scientists:** Faster insights across teams

What is a data product ?

DEFINTION

A data product is a curated, reusable asset that transforms data into actionable insights, designed to meet specific research, scientific, or analytical needs.

PURPOSE

Empowers scientists, researchers, and analysts to efficiently generate insights and drive innovation



Curated, re-usable asset



Transforming Data

INTO



Actionable Insights



Created with specific goal

Data Types in Our Data Product

Clinical

- Demographics
- Survival
- Response
- Tumor stage
- Disease sp. biomarkers

Biomarker

- Gene Expression
- Mutation
- Copy Number Variation
- Tumor Mutation Burden
- BRCA, HPV, MSI status

Metadata

- Sample level
(recurrence, specimen, disease)
- Experimental level
(assay, enrichment, library)

Note: Above list is representative, not exhaustive

Data Harmonization Challenges

Clinical



Manual study-by-study data check



Verify clinical data reflects latest date cutoffs



Non-standard view names lack study milestone and date cutoff. Difficult to trace and identify without an expert



Collaboration trials introduce few variable naming inconsistencies, unit mismatches, and selective data capture for certain parameters

Biomarker



Manual & Expert-Heavy Process to review 300+ views across 20+ trials



Extensive verification needed as one small change alters interpretation



Time intensive effort to fetch missing / hidden metadata information e.g. tumor type (primary, metastatic)



Analysis package specific output layout leads to variations in schema demands extra efforts for harmonization.

Data Harmonization Strategy - Clinical



- ✓ **Review clinical data available in the datalake**
- ✓ **Verify if the data corresponds to latest date cutoffs**
- ✓ **Ingest SDTM / ADaM files as identified**
- ✓ **Defined standardized view naming schema**
<compound>_<studyid>_<adam>_<dataset-name>_<study-milestone>_<cutoff-date>_<dev/prod>
p2026_abc1002_adam_adtte_finaldbl_20200205_prod
- ✓ **Published SDTM / AdaM data with harmonized names in each study**
- ✓ **Defined integrated master view schemas (Base View , Disease – specific View)**
- ✓ **Generated the specification and R code to produce harmonized master views**
- ✓ **Validation checks performed by sIDB matching and spot checks for other key variables**

Harmonization : Master Schema

22 studies, 8000+ subjects, 50+ parameters

Clinical – Base View

Demographics, Treatment Details, Population Flags , Survival & Response variables, Discontinuation Info., Tumor Location & Stage

Clinical – Disease Specific Biomarker View

30 disease sp. biomarkers covering 13 indications captured in clinical SDTM & ADaMs (e.g. HPV, MSI, MCPyV, ALK, BRCA1/2 status)

Data Harmonization Strategy - Biomarker



- ✓ **Generated master list of 300+ biomarker views available**
- ✓ **Classified these into main-group and sub-group categories**
- ✓ **Manually reviewed all views + aligned with experts**
- ✓ **Selected key columns; identified harmonization gaps**
- ✓ **Designed the master view schema and its specification based on review**
- ✓ **Dynamic R code to produce harmonized master views**
- ✓ **Technical & Biological QC validation plans in progress**

Decision Tracker

Key decisions tracked documenting rationale, owners, and timing for full transparency.

Harmonization : Master Schema

15 studies, 6 key biomarkers, 50+ biomarker views used, 3000 + samples across biomarker types

Mutation Master View

Subject ID, Sample ID, Chromosome number and location, Reference and Alternate nucleotide, Read depth, Allele Fractions, Gene Symbols, Variant Type & effect, Exon number

RNASeq Master View

Subject ID, Sample ID, Analysis Workflow details, Gene ID, Chromosome position, Quant Feature, Counts, TPM, Experimental metadata, Genome Version and Build

Team : Oncology Research Unit

Question : Assess Target X in specific cancer indication for specific conditions to support an ongoing hypothesis in early target discovery

Conditions : Assess relation of clinical features (OS, PFS, cBOR) with

1. Target gene expression
2. Evidence of links to EMT signatures
3. Deletions at a specific chromosomal location of interest

Hypothetical scenario (Pre-harmonization) : Efforts would be needed to ensure we have latest data + invest time (few weeks) and efforts to generate view with all desired information, prior beginning analysis

Current scenario (Post-harmonization) : Harmonized master view available. This was directly used to pull values and support quick analysis (few days) for this request delivering insights to the target discovery team for their hypothesis analysis

Key Takeaways



Every dataset is an opportunity

Turn data into solutions for everyone to use.



Easy access via unified data product

Harmonized data in one place speeds up work.



Plan templates + strategy upfront

Team up with experts to shape integrated views.



Invest time now & replicate in future

Blueprint enables and accelerates future products.

THANK YOU

Special thanks for contribution & support in the project
- Olga Bogatyrova, Qi Liu, Alex Rolfe, Eike Staub & team

