**Paper DS05**

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

Namrata Padmashali, Merck Specialities Pvt Ltd, Bengaluru, India

## ABSTRACT

Have you ever considered how a data product could empower R&D functions such as Research, Development and Medical Units, by delivering integrated insights? For compound P2026*, it enabled identification of clinically relevant bladder cancer subtypes through survival data from clinical trials, expression analysis for emerging Antibody-Drug Conjugate (ADC) therapies, and immune landscape profiling. This paper details the Unified Meta-Study Data Product (Meta-Clinical initiative), harmonizing 22 clinical trials' clinical, biomarker, genomics, and metadata into analysis-ready views like clinical_base, clinical_disease_biomarker, mutation_master, and rnaseq_master. Inspired by The Cancer Genome Atlas (TCGA), it creates a scalable "single source of truth", overcoming fragmentation and expertise gaps. Key outcomes include harmonized datasets for lifecycle management, trial design, biomarker selection, target identification, and decision-making, streamlining data preparation from months to days for diverse users including data analysts, clinicians, and biological scientists.

** P2026 is a hypothetical name substituted for our actual asset to maintain confidentiality*

## INTRODUCTION

Pharmaceutical organizations maintain GxP-regulated datalakes housing comprehensive data such as clinical trials, genomics, biomarkers, proteomics, and digital pathology. These datalakes are typically organized studywise with one folder per trial to support trial-specific activities. Each trial's data may be in heterogeneous file formats: sas7bdat / xpt / csv for SDTM and ADaM datasets, fastq / bam / vcf for genomics, and csv / xlsx / pdf for biomarker analyses. In Merck's data lake, tabular data is published as 'views' for downstream querying, analytics, and inclusion in self-service apps. Queryable views abstract file access, enabling R/self-service applications.

The combined 22 trials of data for P2026 represent a treasure trove for clinical and pre-clinical scientists. R&D scientists pursue cross-domain questions such as target discovery, drug repositioning, subtype analyses - but face a harsh reality. Non-experts in statistical programming, biostatistics, or bioinformatics spend weeks/months gathering data across studies, harmonizing messy variables, and preparing for analysis.

The Meta-Clinical initiative delivers a Unified Meta-Study Data Product: analysis-ready master views transforming fragmented datalakes into insights. This paper outlines the "data hunt" problem, data product definition, harmonization – challenges, strategy, master schema, and scalable R&D blueprint.

### WHAT IS A DATA PRODUCT ?

A data product represents a curated, reusable asset that transforms raw, fragmented data into actionable, analysis-ready insights designed for multiple research needs. Unlike standard Clinical Study Reports (CSRs) or regulatory submissions focused on GxP compliance, data products prioritize flexibility for non-regulated research questions-delivering integrated views across clinical outcomes, biomarkers, and biological metadata for diverse scientific needs.

In our scenario, Unified Meta-Study Data Product transforms clinical, biomarker and genomics data for supporting actionable insights such as target evaluation, drug repositioning analysis, subtype analysis and more.
A Data Product could be for one or more compounds as per design.

### WHAT IS HARMONIZATION ?

Data harmonization systematically standardizes heterogeneous variables, structures, and metadata across studies into consistent, queryable master views. Beyond simple database joins, true harmonization incorporates variable standardization, multi-column aggregations, and cross-domain linkages. The outcome: researchers query one or two harmonized master tables instead of navigating 22 P2026 trial folders or 40+ views, enabling analysis in hours versus months of preparation.

## HARMONIZATION CHALLENGES

| Challenge | Clinical Data | Biomarker Data |
|---|---|---|
| Fragmentation | Navigate 22 study folders and verify latest date cutoffs by manually verifying datasets | Manual and expert- heavy review for 300 biomarker trials |
| Messy Variables | Differing variables names across different studies. Legacy or collaboration trials add more to the variations<br>eg. ECOGBL vs BLECOG for ecog baseline status | Variable name differs by assay type or vendor for same information captured across studies<br>eg. pos.chrom vs pos_chrom |
| Non-Standard View Names | p2026_clinical_adam_adsl_dev (no milestone/cutoff information) | - |

All these challenges add further to time delay.

## HARMONIZATION STRATEGY

**Clinical Strategy: Review → Ingest → Standardize → Design → Validate**

- **Review**: Clinical experts assessed 22 folders for data availability & cutoffs
- **Ingest**: Ingested SDTM/ADaM files per study as required
- **Standardize View Naming:** Defined standard view naming schema to overcome challenges in view names.
  Template : <compound>_<studyid>_<adam/sdtm>_<dataset>_<milestone>_<cutoff>_<dev/prod>
  *Example* : p2026_abc1002_adam_adtte_finaldbl_20200205_prod
- **Republish**: All views republished per study
- **Master Design**: Specification → R code → *clinical_base* + *clinical_disease_biomarker*
- **Validate**: sIDB (Study Integration Database) matching + spot checks for other key variables

**Biomarker Strategy: Classify → Review → Design → Harmonize → Validate**

- **Classify**: Cataloged 300+ biomarker views into main/sub-group categories (biomarker-HLA, biomarker-MSI, mutation-CNV, mutation-SNP, etc.)
- **Review**: Manual inspection + domain expert alignment to select critical columns and identify harmonization gaps
- **Design**: Master view schema and the relevant specifications
- **Harmonize**: Dynamic R code standardizing gene IDs (Ensembl/NCBI), units (TPM), and experimental metadata
- **Validate**: Technical + biological QC validation ongoing

## DECISION TRACKER

Various decisions taken throughout the project are documented in the tracker with rationale, owners, timing for transparency and traceability.

## HARMONIZED MASTER VIEW SCHEMA

| Master View | Scope | Key Contents |
|---|---|---|
| **clinical_base** | Subject-level essentials | Demographics, survival (OS/PFS), response (cBOR), tumor stage/location, treatment details |
| **clinical_disease_biomarker** | Disease-specific markers | HPV, ALK, CA-125, PD-L1, MSI, BRCA1/2, Fuhrman grade (~30 markers) |
| **mutation_master** | Genomic variants | Subject/Sample ID, chr:pos, ref/alt, allele fraction, gene symbol, variant effect, TMB |
| **rnaseq_master** | Expression data | Gene ID, TPM, fusions, signatures, workflow, genome build, experimental metadata |

## CONCLUSION

The Unified Meta-Study Data Product transforms GxP datalake chaos into scalable, compound-specific data asset. By resolving fragmentation through systematic harmonization, P2026 researchers across R&D functions now access cross-domain insights instantly rather than after months of data hunting. This blueprint extends beyond single compounds - future iterations could incorporate additional assets.

## ACKNOWLEDGMENTS

Thanks to Olga Bogatyrova, Qi Liu Alex Rolfe, Eike Staub & the wider team

## CONTACT INFORMATION

Author Name: Namrata Padmashali
Company: Merck Specialities Pvt Ltd
Address: Bengaluru, India
Email: namrata.padmashali@merckgroup.com
Linkedin: https://www.linkedin.com/in/namratapadmashali/