# A UNIVERSAL DATA MODEL (UDM) FOR LONGITUDINAL INTEGRATION OF DISPARATE BIOMARKER AND IN-LIFE PATIENT DATA AUGMENTED BY MACHINE LEARNING

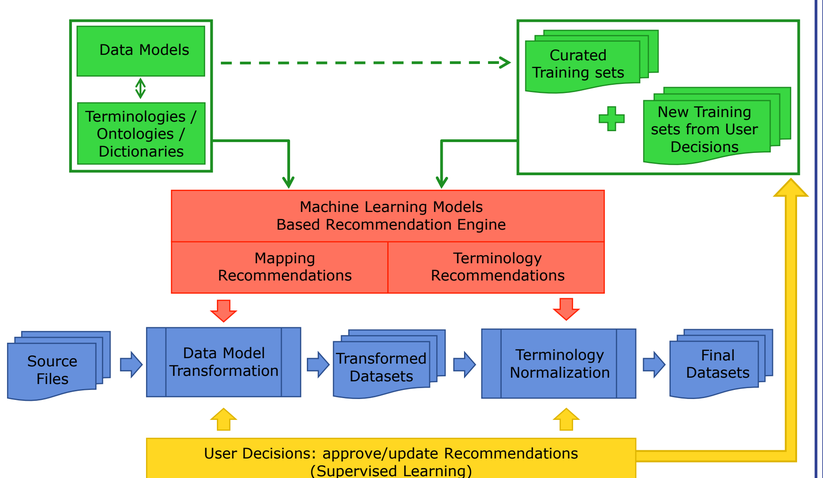
## Introduction

Translational and precision medicine development in immuno and gene therapies rely on biomarker data from assays including genomics, proteomics, IHC, Flow cytometry and cell phenotyping data. Biomarkers are not only generated from the patient biosamples, but also from the biomanufacturing sites such as in adoptive immune cell biomanufacturing for novel immunotherapies. Extracting valuable insights from these disparate data sources and integrating it to clinical data to relate to patient outcome and/or discover and validate relevant biomarkers are met with challenges of ingestion, harmonization and integration of disparate data with the clinical data. Key decisions and ideas that impact study design of future clinical trials depend on gaining insights rapidly from such integrated data on patients or stratified cohorts. Systematic curation with self-validation for completeness and consistency is time consuming and difficult without technology. Xbiom, built on Machine Learning and Universal Data Modelling effectively solves these challenges of disparate, big and varied data sources. Xbiom's Smart Transformation platform, can make ingestion, curation and harmonization process automatic and is also capable of processing both streamed data as well as in batch mode.

## Data Transformation Workflow

Xbiom's Machine Learning algorithms work on the principle of supervised learning. First, for the given data set Recommendation Engine recommends the Target model along with referencing to external registries for Data transformation. It is then followed by Mapping recommendations by machine learning models based on the trained datasets. The Xbiom platform is agile by having the ability to retrain the algorithm for the newer/unseen type of data types and structures but also robust in predicting the column for a recommendation in manual approval in case of unseen data types. Machine Recommendations can be further approved by the users which will be considered as learning in the next iteration.

Mapped data is further enriched by the Terminology Recommendation, wherein, for instance for protein names in Biomarker data will be enriched with species-specific Harmonization of proteins with reference to the Global registry UniProt. Similarly, Harmonization of Genes, cell-expressed proteins, cytokines etc., will be done with respect to the referenced registries. The curated and transformed output is quality checked and which can be viewed in the QC dashboard for the quality process.




*Workflow of Data Transformation*

## Smart Transformation and Harmonization

Raw Input File

| LLOQ | D | M | Y | STUDYID | STAT | VENDOR | SUBJIDN | TPT | TESTCD | ORRES | ORRESU | METHOD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.1 | 23 | 4 | 2019 | PC123 | | CROSSTICS | 1001 | C1D1 | IL_6 | 6.6 | pg/mL | multiplex immunoassay |
| 1.1 | 23 | 4 | 2019 | PC123 | | CROSSTICS | 1001 | C1D1 | VEGF_A | 68 | pg/mL | multiplex immunoassay |
| 1.1 | 23 | 4 | 2019 | PC123 | | CROSSTICS | 1001 | C1D1 | VEGF_C | 1432 | pg/mL | multiplex immunoassay |
| 1.1 | 23 | 4 | 2019 | PC123 | | CROSSTICS | 1001 | C1D1 | VEGF_D | BLQ | pg/mL | multiplex immunoassay |
| 1.1 | 23 | 4 | 2019 | PC123 | | CROSSTICS | 1001 | C1D1 | TMEM16B | 12.5 | pg/mL | multiplex immunoassay |
| 1.1 | 23 | 4 | 2019 | PC123 | | CROSSTICS | 1001 | C1D1 | DPPI | 23 | pg/mL | multiplex immunoassay |
| 1.1 | 23 | 4 | 2019 | PC123 | | CROSSTICS | 1001 | C1D1 | CYPIIA6 | 15.7 | pg/mL | multiplex immunoassay |
| 1.1 | 23 | 4 | 2019 | PC123 | | CROSSTICS | 1001 | C1D1 | Connexdenn 2 | 1234 | pg/mL | multiplex immunoassay |
| 1.1 | 23 | 4 | 2019 | PC123 | | CROSSTICS | 1001 | C1D1 | FMR2P | 112.34 | pg/mL | multiplex immunoassay |
| 1.1 | 23 | 4 | 2019 | PC123 | | CROSSTICS | 1001 | C1D1 | AFAP-110 | 11.5 | pg/mL | multiplex immunoassay |

Disparate Data - Sample formats for Immunoassay data received from Vendors

| STUDYID | SUBJIDN | SPID | TPT | TESTCD | ORRES | ORRESU | LLOQ | STAT | REASND | D | M | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PC123 | 1001 | 501 | C1D1 | IL_6 | 6.6 | pg/mL | | 1.1 | | 23 | 4 | 2019 |
| PC123 | 1001 | 501 | C1D1 | VEGF_A | 68 | pg/mL | | 1.1 | | 23 | 4 | 2019 |
| PC123 | 1001 | 501 | C1D1 | VEGF_C | 1432 | pg/mL | | 1.1 | | 23 | 4 | 2019 |
| PC123 | 1001 | 501 | C1D1 | VEGF_D | BLQ | pg/mL | | 1.1 | | 23 | 4 | 2019 |

Machine Readable Columnar Format

| Study | Order | TestProfile | SampleName | Species | SampleType | Anticoagulant | Analyte | Units | LOD | LLOQ | LowRange | HighRange | Result | Comment |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PC001 | P002 | MAP | PC123_Screen | Human | Plasma | EDTA | 6Ckine | pg/mL | 3.6 | 11 | 357 | 1330 | 410 | |
| PC001 | P002 | MAP | PC123_Screen | Human | Plasma | EDTA | Adiponectin | ug/mL | 0.02 | 0.13 | 1.8 | 15 | 3.8 | |

| Sample | Assay | Well | Spot | Dilution | Concentration | Sign al | Mean | Calc. Concentration | Calc. Conc. CV | Calc. Conc. Mean | SampleID | SUBJID | Compound Name | X250 | X5 | Sample type |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PC001 | IL-6 | A03 | 7.00 | 1.00 | | 659.00 | | | | | | | | | | |
| | | | | | | 536.00 | | 598.00 | 0.00 | 52.38 | 0.00 Biopsysamp1 | PC001-121 | Vehicle | | | Biopsy |
| PC001 | GM-CSF | A04 | 7.00 | 1.00 | | | | | | | | | | | | |
| | | | | | | 00 | | 598.00 | 0.00 | 52.38 | 0.00 Biopsysamp2 | PC001-121 | Vehicle | | | Biopsy |

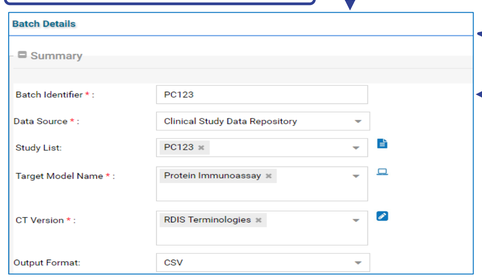
| Unique Subject Identifier | Lab Test or Examination Name | Category for Lab Test | Result or Findings as Collected | Unit of the Original Result | Specimen Condition | Method of Test or Examination |
|---|---|---|---|---|---|---|
| PC11801 | IL-6 | WB | | 0.6 pg/mL | | Human whole Blood Assay |
| PC11801 | IL-8 | WB | | 282.8 pg/mL | | Human whole Blood Assay |

| Analytes | | | | Granulocyte-Macrophage Colony-Stimulating Factor (GM-CSF) | Interleukin-2 (IL-2) |
|---|---|---|---|---|---|
| Units | | | | pg/mL | pg/mL |
| LOD | | | | 15 | 82 |
| LLOQ | | | | 29 | 142 |
| Low Range | | | | | |
| High Range | | | | ND | ND |

| Samples | SiteID | SubjectID | SampleID | Visit Type | Date of Collection | Time of Collection | | |
|---|---|---|---|---|---|---|---|---|
| PC00157763284 | A001 | PC123 | 54323466 | C1D1 | 09/17/2020 | | 43 | 356 |

Human Readable Pivoted Format

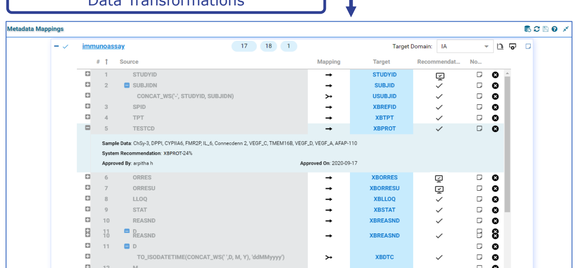
Target Model for Transformation




Source files transformation to a selected Target model along with reference to external registries for Data harmonization

| Variable Label | Value Coded By | Controlled Terms,Format... | Core | Data Type | Notes |
|---|---|---|---|---|---|
| Study Identifier | | | Required | Text | Study Identifier |
| Domain | | | | Text | Domain Name |
| Subject Identifier for the Study | | | Required | Text | Subject Identifier |
| Unique Subject Identifier | | | | Text | Identifier used to uniquely identify a subject ac... |
| Secret Flag | | | | Integer | To mark Secret flag for Subjects (0-Restricted,... |
| Specimen Type | Controlled Terms | SPECTYPE | | Text | Sample Type or Matrix or Tissue type |
| Ontology ID for Specimen Type | | | | Text | Ontology ID for Sample type |
| Category of Specimen Type | | | | Text | Sample Category (Ex: BODY LIQUIDS) |
| Source Visit Name | | | | Text | Visit Name |
| Visit Name | | | | Text | Visit Name. Needs to map to visit name from ... |
| Visit Number | | | | Float | Visit Number. Needs to map to visit name fro... |
| Protein Name | External Dictionary | UniProt | Required | Text | Source Protein Name |
| Source Protein ID | | | | Text | Protein ID |
| Gene Symbol | External Dictionary | NCBI Gene Info | | Text | Gene Symbol |
| Source Gene ID | | | | Text | Source Gene Identifier NOT identical to NCBI ID |
| Result or Finding in Original Units | | | | Text | Results or Findings |
| Original Units | | | | Text | Units of the Result collected |

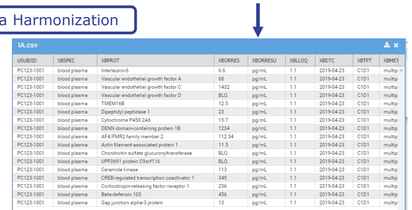
Data Transformations



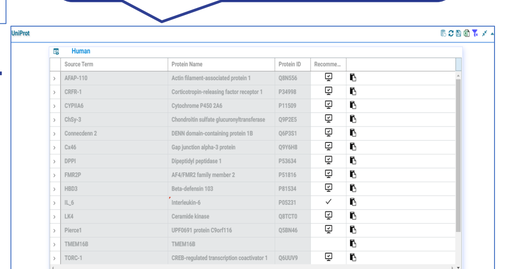

Mapping of Source data to Target model variables, Mapping Recommendations by machine learning models based on the Training datasets & Users previous decision. Machine Recommendations can be further approved by the user which will be considered as Learning in the next iterations

Species specific Harmonization of Proteins with the Global registry UniProt. Recommendation of the Protein names based on look-up of source protein names in the registries.

Data Harmonization



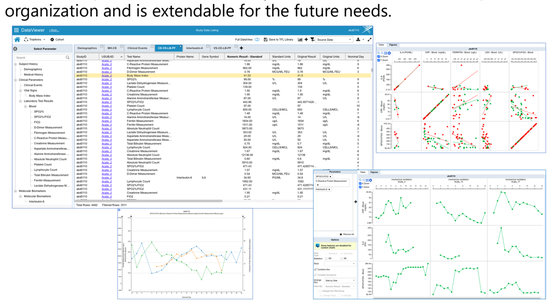

Standardized Output File




## Longitudinal Integration of Biomarker Data & patient data facilitated by UDM for visualization and analysis

Curated and harmonized data is stored in a Universal Data Model (UDM) which holds study and assay data of subjects or samples and their attributes in a simple, indexable form. The indexation facilitates instant search through query masks to yield highly stratified cohorts and data. Search and detection of signals is enabled by the integration of assay data into longitudinal patient data.

As shown in the immersive graphics below on the left sidebar listing the available data for the searched cohort enables handpicking of Biomarker which can be correlated across the cohort with another biomarker(IHC, TMB, FACS, Nanostring etc.,) or with the clinical data (adverse event, clinical endpoints, Medical history etc.,). UDM can be unique to the organization and is extendable for the future needs.



## Conclusion

Challenges in translational medicine and other research in drug development involving different biomarkers are the difficulty in getting the cleaned biomarker data which is integrated with the clinical data to ask questions to gain insights into the study. Manual curation and harmonization can be time-consuming and would be challenging with steamed data in the ongoing trials.

Automated curation, standardization and harmonization by the Machine Learning platform in Xbiom is quality controlled and can process the streamed data. Data in the UDM is ready for the statistical analysis and enables the monitoring of ongoing trials for early insights and decision making. Harmonized data in the UDM can also act as a training data set in machine learning designed for biomarker discovery and other insights generation as in precision and translational medicine. The data governance system in Xbiom's allowing for the role-based access can be accessed safely in the browser as well as through the API.
Quick and timely decision enabling process can cut down the time and cost in drug development.


POINT CROSS Life Sciences

Poster By:
Dr. Rajalaxmi S,
Arpitha Hanumanthiah

Contact: get_started@pointcross.com