Smart Transformation of Clinical & Nonclinical Data for Insights

Machine Learning and AI augmented transformation of Clinical and Biomarker Data for semantically harmonized access

PhUSE Connect 2019
Smart Transformation

➢ Problem: Disparate data

✓ Inconsistent data models and terminologies
✓ Unstructured descriptions from studies
✓ Challenging to search across studies

➢ Solution: Smart Transformation

✓ Artificial intelligence and machine learning augmented transformation of raw as-collected data from data lakes into standardized, usable, accessible, searchable repositories
Smart Transformation

➢ Purpose

☑ Curate by looking up Data Lake through Search
☑ Transformation Facility to automatically identify re-usable data from raw, legacy or incoming sources.
☑ Conversion recommendations of selected data from batches of studies into Structured/Standard format for searchable repository

➢ Benefits for Users

☑ AI machine learning algorithms use Training sets (Data) & Terminology Governance to recommend mappings
☑ “Exception based decisions” for Data Managers to save time
☑ Data Validation and QC dashboards to review Data Consistency, Data Formats & Terminology issues
Machine Learning Algorithms

- Artificial Neural Networks based Machine Learning Algorithms

- Pre-trained with Clinical, Non-Clinical and Biomarkers data, and continually learns based on user applied mapping and transformation decisions

- Recommendation algorithms for data mapping, transformation functions and data pivots.

- Recommendation algorithms for Terminology Normalization based on many industry standard databases – CDISC CT, MedDRA, WHODD, GeneInfo, UniProt, miRBase,...
Artificial Neural Networks for Smart Transformation
Smart Transformation Process Flow

Data Models
Terminologies / Ontologies / Dictionaries
Curated Training sets
New Training sets from User Decisions

Machine Learning Models based Recommendation Engine
Mapping Recommendations
Terminology Recommendations

Source files
Data Model Transformation
Transformed Datasets
Terminology Normalization
Final Datasets

User Decisions: approve/update the Recommendations
Mapping Recommendation (Target Domains & Variables)

Source file metadata

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTNO</td>
<td>Patient Number</td>
</tr>
<tr>
<td>SEX</td>
<td>Sex</td>
</tr>
<tr>
<td>DOB</td>
<td>Date of Birth</td>
</tr>
<tr>
<td>INVSITE</td>
<td>Site</td>
</tr>
</tbody>
</table>

Recommended mappings

- PTNO → DM.SUBJID
- SEX → DM.SEX
- DOB → DM.BRTHDTC
- INVSITE → DM.SITEID

Source data file

<table>
<thead>
<tr>
<th>PARAM</th>
<th>MATRIX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basophils</td>
<td>Blood</td>
</tr>
<tr>
<td>Eosinophils</td>
<td>Blood</td>
</tr>
<tr>
<td>Leukocytes</td>
<td>Blood</td>
</tr>
<tr>
<td>Glucose</td>
<td>Urine</td>
</tr>
</tbody>
</table>

Recommended mappings

- PARAM → LB.LBTEST
- MATRIX → LB.LBSPEC
Terminology Recommendation (Target Terms)

Transformed dataset

<table>
<thead>
<tr>
<th>LBSPEC</th>
<th>LBTEST</th>
<th>LBORRESU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urine</td>
<td>CRE</td>
<td></td>
</tr>
<tr>
<td>Blood</td>
<td>WBC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UCREA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UGLUCR</td>
<td></td>
</tr>
<tr>
<td>Blood</td>
<td>BASO</td>
<td>G/L</td>
</tr>
<tr>
<td>Blood</td>
<td>BASO</td>
<td>%</td>
</tr>
</tbody>
</table>

Recommended Terms

<table>
<thead>
<tr>
<th>LBSPEC</th>
<th>LBTESTCD</th>
<th>LBTEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urine</td>
<td>CREAT</td>
<td>Creatinine</td>
</tr>
<tr>
<td>Blood</td>
<td>WBC</td>
<td>Leukocytes</td>
</tr>
<tr>
<td>Urine</td>
<td>CREAT</td>
<td>Creatinine</td>
</tr>
<tr>
<td>Urine</td>
<td>GLUCRT</td>
<td>Glucose/Creatinine</td>
</tr>
<tr>
<td>Blood</td>
<td>BASO</td>
<td>Basophils</td>
</tr>
<tr>
<td>Blood</td>
<td>BASOLE</td>
<td>Basophils/Leukocytes</td>
</tr>
</tbody>
</table>
Smart Transformation

- Recommendation Engine
  - Created using TensorFlow™ and Keras™
  - Models trained to classify data domains and variables
  - Classification results are summarized with a confidence score
  - User reviews results and can change if necessary

- Continuous learning
  - User changes tracked
  - Training datasets updated to capture user input

- Automation to reduce time and effort
  - Identifying needed data
  - Semantically mapping and transforming to target format
SUPPLEMENTARY SLIDES
Smart Transformation – High level steps

Start
- Create a Batch
  - Select,
    - Source
    - Studies
    - Target Model
    - Applicable CTs
  - Save & Initiate Process

Transformation
- Dataset Transformations
- Metadata Mappings
- Terminology Normalization
- Normalization to External dictionaries
- Additional Transformations

Validation
- Preview in each stage
- Validation and QC Check on generated Outputs
- View Output files
- Store output files in destination

End
Source file “basecond” data is recommended to “Medical History (MH)” with high confidence. Source columns CPEVENT & ACTEVENT are recommended to VISIT values. Based on the data, CDMX column is recommended as MHTERM & AETERM.
### Terminology Recommendation (Target Terms)

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>Expression</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activated Partial Thrombin</td>
<td>Partial Thromboplastin Time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Albumin</td>
<td>Albumin Measurement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alkaline Phosphatase</td>
<td>Alkaline Phosphatase Measurement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacteria</td>
<td>Bacterial Count</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilirubin</td>
<td>Total Bilirubin Measurement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blood</td>
<td>Blood</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Casts</td>
<td>Cast Present Or Absent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crystals</td>
<td>Crystal Present Or Absent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epithelial Cells</td>
<td>Epithelial Cell Count</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glomerular Cells</td>
<td>Glomerular Cell Measurement</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Xbiom – Technology Stack

Web Server

RDBMS

Bigdata Processing

Normalized data big table

Data Analysis, Interaction and Search

- Single Sign-on Web Services
- HTTPS

- User Accounts & Roles
- Object meta-data
- Crawler configurations
- And Mappers

- Smart Curation with Machine Learning Algorithms for Data Standardization and normalization.

- Normalized data with Precompiled Summaries and statistics

- Data visualization, Search and interactivity

Orchestra™
**Xbiom - Technical Architecture**

- **API Clients**
  - R, Analytics and Business Intelligence applications
- **Browser Based User Interfaces & Visualizations (IGO)**
- **Web Server**
- **OData REST API**
- **REST Web Services**
- **Document & Metadata Search**
- **Patient Cohort and Biomarker Search**
- **Smart Curation & Transformation Data Ingestion**
  - Mapping Transformation to UDM, Terminology normalization, Data Processing and Indexing
- **Data Models & Metadata Management**
- **Ontology & Terminology Services**
- **Data Crawlers and Integrations**
- **Study & Data Management**
- **Xbiom – Orchestra Platform Features**
  - Storage, Workflows and Collaboration

**Authentication & Authorization Management**

- **R API Client application package as a convenience query wrapper on OData API for Search and query/extract data for a Saved Patient Cohort**

- **Controls for users to Search, View patient data and Save as Cohorts. Native capabilities to pivot the data and visualize graphically through IGOs.**

- **REST Web Services for Application User Interfaces, OData based REST API for data access through third-party analytics and visualization applications**

- **Documents / raw data file Search based on meta-data attributes and content (configurable) from multiple data sources. Federated Search is also possible.**

- **Tiered query mask configuration for Search for identifying/finding patient cohort of interest by scientific user community through many clinical, genotype, phenotype markers.**

- **Data transformations using Smart Curation Machine learning algorithms, and terminology normalization based on configured ontologies/databases, pre-computed references/summaries and Indexing for Search**

- **Define metadata models (Simplified Data Models-SDM) on expected data formats for data governance and data standards management. CDISC meta and SDTM/AdAm IG models are built-in.**

- **Interfaces to Ingest Sponsor Specific Ontologies, Industry standard clinical coding databases (e.g. MedDRA, WHODD, . . .), External Biomarker databases (GeneInfo, UniProt, miRBase, COSMIC, GeneOntology, . . .), Built-in CDISC Controlled Terminologies**

- **Collates Study Data & Metadata from different data sources and systems to create single study view, study folder template configurations, and reprocess controls.**

- **Data crawling interfaces to Biomarker Data Archive, Network shares (Sample availability, clinical data), SharePoint and other third party repositories/APIs.**

- **Single Sign-on Authentication with Active Directory, Roles Based Access Control (RBAC), Workflow automation and collaboration, Metadata storage interfaces to Oracle database, and Data/Document storage into HDFS**

*All product names, logos, and brands are property of their respective owners. All company, product and service names used in here are for identification purposes only. Use of these names, logos, and brands does not imply endorsement.*
THANK YOU

Isaac Mativo < isaac@pointcross.com >
get_started@pointcross.com

PointCross Life Sciences
www.pointcrosslifesciences.com