KNIME Data Talks

Clinical Analysis Dataset Derivation using Visual Programming with KNIME

Robert Adams 20220601

Oncology Digitalization & Computational Science
SBU Oncology
Background – Clinical Data Derivation
A primer into clinical programming and regulatory standards

• Regulatory authorities such as FDA (U.S. Food and Drug Administration) or EMA (European Medicines Agency) require pharma companies to submit their study data in certain standards

• Pharma companies or CROs (clinical research organisations) create study results programmatically usually referred to as “TLFs” (tables, listings, figures)

• A study thereby consists of different derivation levels
  • Usual data flow: (e)CRF / EDC → RAW → SDTM → ADaM → CSR → submission
  • The process is more complex by iterations of data cleaning steps and multiple additional standards and guidelines that need to be considered

(e)CRF / EDC – (electronic) Case Report Form / Electronic Data Capture; SDTM – Study Data Tabulation Model; ADaM – Analysis Data Model; CSR – Clinical Study Report
Background – CDISC (Clinical Data Interchange Standards Consortium)

SDO (standards developing organization)

“enable information system interoperability to improve medical research and related areas of healthcare”

https://www.cdisc.org/standards
https://en.wikipedia.org/wiki/Clinical_Data_Interchange_Standards_Consortium
Background – ADaM (analysis data sets) & Domains

The CDISC Glossary defines these terms as follows:

- Domain: A collection of logically related observations with a common, specific topic that are normally collected for all subjects in a clinical investigation.

ADaM defines dataset and metadata standards that support:

- efficient generation, replication, and review of clinical trial statistical analyses, and
- traceability among analysis results, analysis data, and data represented in the Study Data Tabulation Model (SDTM).

ADaM is one of the required standards for data submission to FDA (U.S.) and PMDA (Japan).

Details on the requirements for FDA are specified in the FDAs Data Standards Catalog for NDA, ANDA, and certain BLA the FDA Guidance on Standardized Data.

Details on the requirements for PMDA can be found on the Advanced Review with Electronic Data Promotion Group p

https://www.cdisc.org/standards/foundational/adam
https://www.cdisc.org/kb/articles/domain-vs-dataset-whats-difference
Background – Clinical Programming & Challenges

- SAS is a de-facto standard proprietary commercial programming language used by the pharmaceutical industry / regulatory agencies
- TLFs are programmed by SAS experts
- Implementation of CDISC (and other standards) using specific languages follow „SOPs“ (standard operation procedures in companies)
  - Leads to highly interconnected dependencies that makes it challenging to try deviating approaches
- Not all functions involved (e.g. DM – data management – responsible for „clean data“) are necessarily trained programmers
- Although highly standardized, every study is different
  - Maybe 80 % is standardized? 20 % need to be adopted from study to study
  - Especially efficacy domains – different end points defined by (complex) study designs
- Industry dependency for certain software providers
  - No realistic chance in near future to use KNIME for submissions to authorities
  - Currently more of a case study (feasibility) to prove that other technical solutions are possible (spark ideas!) and KNIME could be used for non-regulated (internal) work with clinical data
Rationale – Why Visual Programming?

An alternative way of deriving clinical analysis data sets

- Double programming / validation alternative
- Intuitive visualization of data flows within the “program”
  - “The program is the documentation”
- Accessibility to new learners
- Standardization
- “Lowest common denominator”
  - Data science concepts shared across different programming & skills backgrounds (e.g. programmers and DM)

- Immediate clarity of algorithmic approach for anyone with programming / data science backgrounds
- Optimization of workflows straight forward
SAS vs. KNIME

Data flows

Example: Batch categorization with Column Expressions

Meta node (comparable to SAS Macros)

Configured node (not executed)

Commonly required ETL functionality

<table>
<thead>
<tr>
<th>Type</th>
<th>Collection</th>
<th>Replace Column</th>
<th>Output Column</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>DURDIS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>DURDS...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example: Batch categorization with Column Expressions

```
# if (column("DURDIS") >= 12) {
  DURSOR1 = ">=12"
} else {
  DURSOR1 = "<12"

77 => 12
33 => 12
19 => 12
11 => 12
73 < 12
48 => 12
24 => 12
```
ADaM Derivation – ADSL

(subject listings)

Data retrieval

Derivation

Testing & XPT output
ADSL – Data Retrieval

https://github.com/phuse-org/PODR

• “PODR integrates health-related Open Data across agencies”
• Sample data sets
  • SDTM, ADaM … and more
ADSL – Data Retrieval

https://github.com/phuse-org/PODR
ADSL – Derivation

Coping with date formats

![Diagram showing date transformations and calculations]

<table>
<thead>
<tr>
<th>SAS date representation</th>
<th>Define SAS origin</th>
<th>Shift to KNIME dates</th>
<th>Duration from TRTSTD* (e.g. in days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>19800</td>
<td>1960-01-01</td>
<td>2014-01-02</td>
<td>44</td>
</tr>
<tr>
<td>19905</td>
<td>1960-01-01</td>
<td>2012-08-05</td>
<td>76</td>
</tr>
<tr>
<td>19401</td>
<td>1960-01-01</td>
<td>2013-07-19</td>
<td>43</td>
</tr>
<tr>
<td>19724</td>
<td>1960-01-01</td>
<td>2014-03-18</td>
<td>55</td>
</tr>
</tbody>
</table>

*TRTSTD = Treatment Start Date

Allow date calculations

...
ADaM Derivation – ADTTE (time to event)

Data retrieval

Derivation

Visualization & Testing & XPT output
Breakout – Survival Analysis

Crucial **efficacy representation** of clinical trials – „did a drug has an anticipated effect?“

- General: area of **statistics designed for modelling TTE** (time-to-event) data („expected duration of time until event occurs); applications in sociology, economics, engineering, biology etc.
  - E.g. failure in mechanical systems, **survival of a population past a certain time**
- In oncological clinical study context: analyse the **time to disease remission, progression or death for cohorts of patients** or compare different treatments within a clinical trial
- Events typically subject to **censoring** (missing / incomplete) for variety of reasons
  - I.e. subjects are „**lost to follow-up**“ or **drop out** of a study for reasons independent of survival
  - Influences statistics as power decreases by censoring (uncertainty increases)
- **Survivor function** (probability to experience and event by given time, e.g. survival probability after 24 months) and **hazard rate** (instantaneous risk of a subject experiencing an event at a given time)
  - Wilcoxon test and log-rank test used to calculated differences between groups (H₀: groups have the same hazard)
  - Usually Kaplan-Meier plots to present survival data are used

\[
S(t) = P(T \geq t)
\]

\[
\hat{S}(t) = \prod_{i=1}^{t \leq t} \left( \frac{n_i - d_i}{n_i} \right)
\]
ADTTE – Derivation
Kaplan-Meier Statistics / Plotting

\[ S(t) = P(T \geq t) \]

\[ \hat{S}(t) = \prod_{i=1}^{t} \left( \frac{n_i - d_i}{n_i} \right) \]

Caution: Details matter!
E.g. CI calculations have subtle changes in different programming languages and might have specific definitions
ADTTE – Derivation

Utilizing R to write XPT output (regulatory authority requirement)
Core of GxP:
Validation, Testing, Logging, Reporting

They weren't so much different, but they had different goals
Double Programming Validation

Dear PODR – no offense 😊 you have a typo!

Gold standard data set; PODR ADTTE (or SAS derived data)

For numeric: min / max (range) differs
For nominal: different unique values

For numeric: min / max (range) differs
For nominal: different unique values
Double Programming Validation

<table>
<thead>
<tr>
<th>Derived data table</th>
<th>Value differences (e.g. rounding errors, or different data types [int vs. double])</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold standard data set: PODR ADTTE (or SAS derived data)</td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HEIGHTBL</td>
<td>135.89</td>
<td>135.9</td>
</tr>
<tr>
<td>WEIGHTBL</td>
<td>55.79</td>
<td>55.8</td>
</tr>
<tr>
<td>EDUCLVL</td>
<td>15</td>
<td>15.0</td>
</tr>
<tr>
<td>DURDIS</td>
<td>44</td>
<td>43.9</td>
</tr>
</tbody>
</table>
Testing

https://www.knime.com/blog/enter-the-era-of-automated-workflow-testing-and-validation

All relevant testing scenarios can be covered
Logging

Export executed workflow summary

Verbose XML document traces all data points / transformations throughout a workflow

Use third party tools or read back into KNIME and use the provided Analyzer component

Environment & Metadata
Logging

Export executed workflow summary

Node details:
- For each node input, state, configuration, settings, variables
- Data at input port, data at output port

More verbose than usual SAS logs
Conclusion & Take-aways

• Powerful (feature complete) alternative for working with (clinical) data

• Clinical data derivation possible
  • Alternative approach for double programming / validation of SAS derivations
  • Verbose logging & validation capabilities

• Visual programming especially for pipeline optimization, automation & educational purposes with great potential

• Highly regulated environment and current standards

• Industry adoption

• Existing pipelines, SOPs and study continuity
Acknowledgements

Seema Nair
Bayer – Onc SBU – Statistical Programming

Anupama Sheoran
Bayer – Onc SBU – Statistical Programming

José C. Lacal
PHUSE – PODR

Cornelia Fulgenzi
Bayer – Onc SBU – Statistical Programming
Contact Information

Robert Adams
Lead Computational Scientist
Bayer – Oncology Digitalization & Computational Science

Clara Beck
Senior Computational Scientist

robert.adams@bayer.com
KNIME Data Talks

Clinical Analysis Dataset Derivation using Visual Programming with KNIME

Main Author: Robert Adams
Co-Author: Clara Beck

Thank you for the attention