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

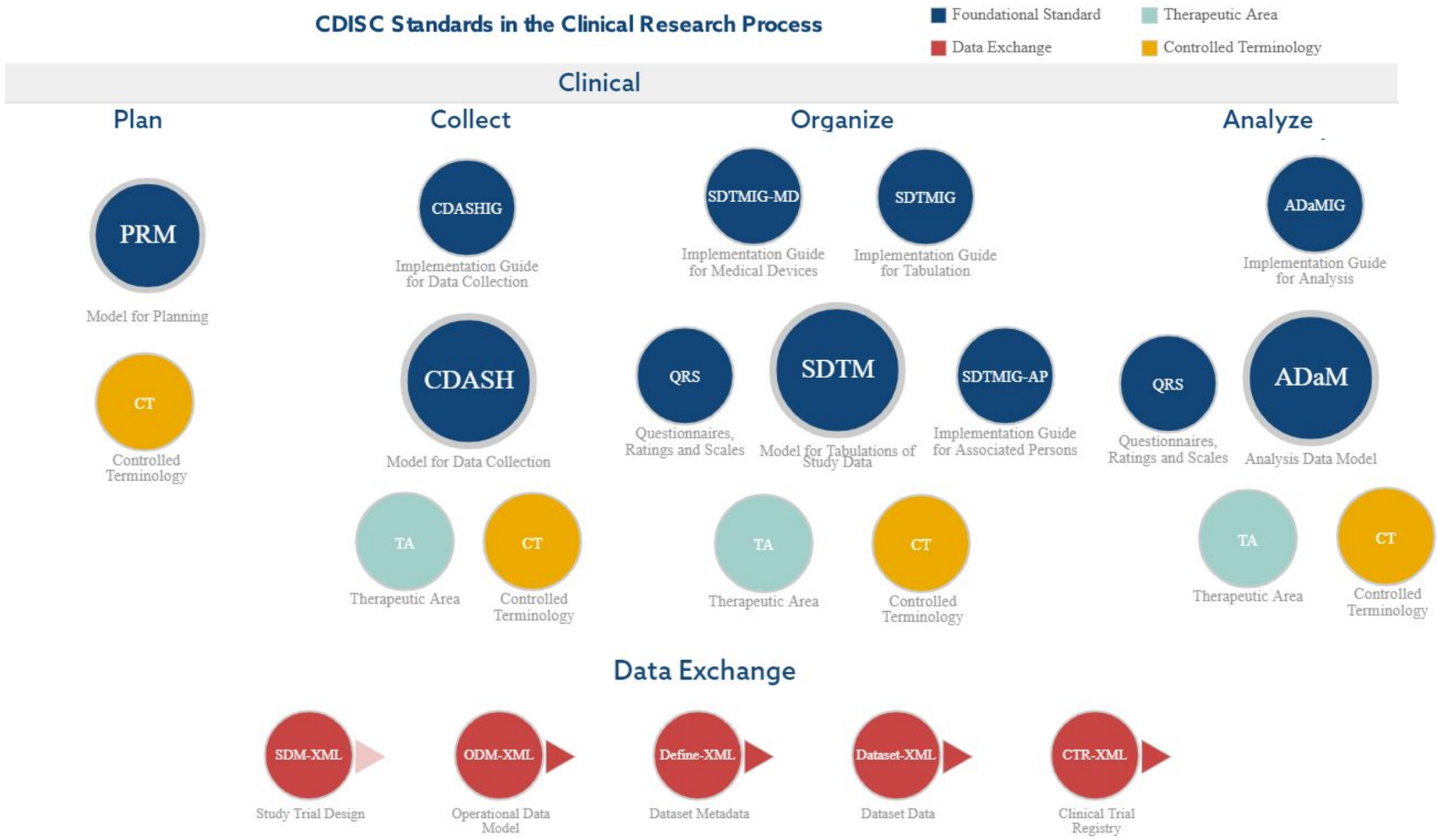
A primer into clinical programming and regulatory standards

- 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 additonal 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

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.

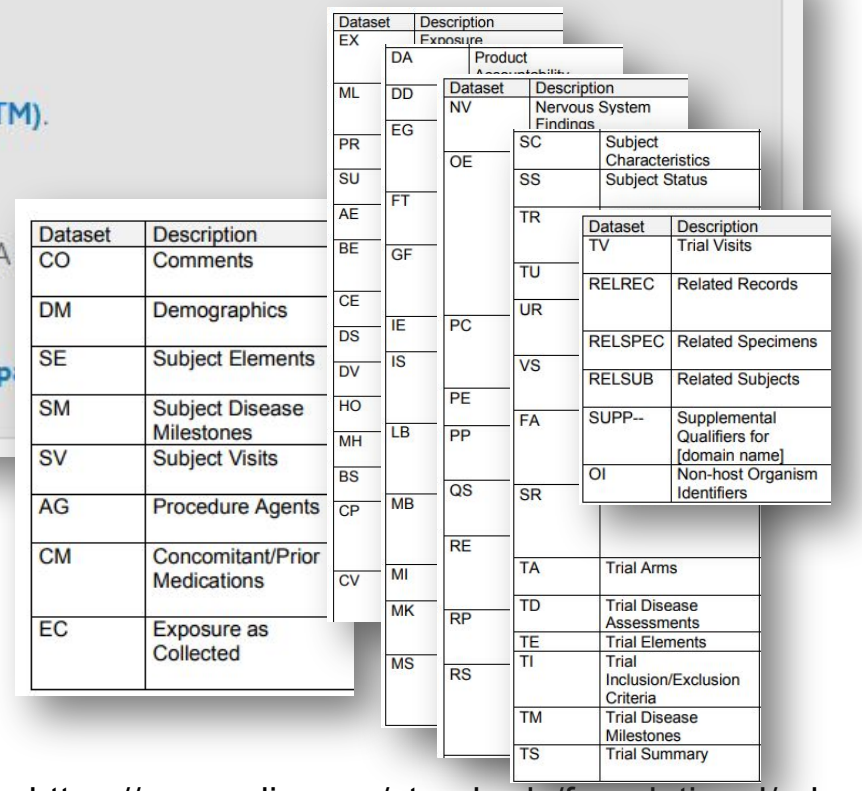
ADaM defines dataset and metadata standards that support:

- 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 **FDA's 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**



Dataset	Description
CO	Comments
DM	Demographics
SE	Subject Elements
SM	Subject Disease Milestones
SV	Subject Visits
AG	Procedure Agents
CM	Concomitant/Prior Medications
EC	Exposure as Collected

Dataset	Description
EX	Exposure
ML	
PR	
SU	
AE	
BE	
CE	
DS	
DV	
HO	
MH	
BS	
CP	
CV	

Dataset	Description
DA	Product
DD	
EG	
FT	
GF	
IE	
IS	
LB	
MB	
MI	
MK	
MS	

Dataset	Description
NV	Nervous System Findings
OE	
PC	
PE	
PP	
QS	
RE	
RP	
RS	

Dataset	Description
SC	Subject Characteristics
SS	Subject Status
TR	
TU	
UR	
VS	
FA	
SR	
TA	Trial Arms
TD	Trial Disease Assessments
TE	Trial Elements
TI	Trial Inclusion/Exclusion Criteria
TM	Trial Disease Milestones
TS	Trial Summary

Dataset	Description
TV	Trial Visits
RELREC	Related Records
RELSPEC	Related Specimens
RELSUB	Related Subjects
SUPP--	Supplemental Qualifiers for [domain name]
OI	Non-host Organism Identifiers

cdisc

Study Data Tabulation Model
Implementation Guide: Human Clinical Trials
Version 3.4 (Final)

https://www.cdisc.org/standards/foundational/adam

https://www.cdisc.org/kb/articles/domain-vs-dataset-whats-difference

Background – Clinical Programming & Challenges

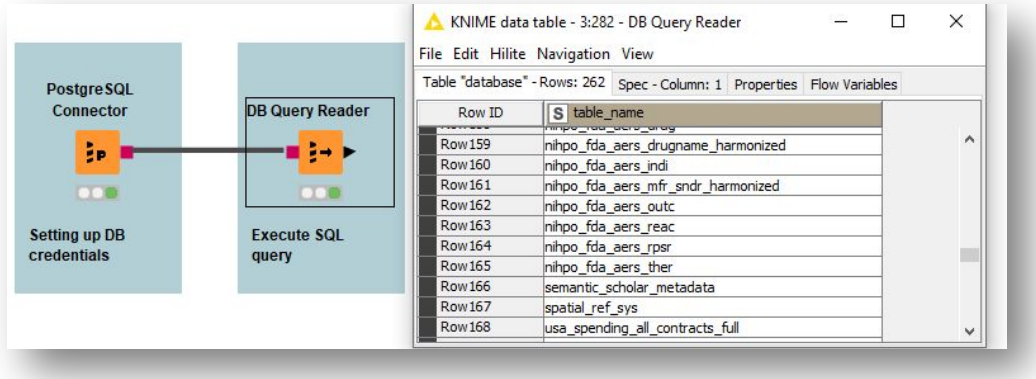
- SAS is a de-facto standard proprietary commercial programming language used by the pharmaceutical industry / regulatory agencies
- TLFs are programmed by SAS experts
- Implementation of CDISC (and other standards) using specific languages follow „SOPs“ (standard operation procedures in companies)
  - 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)



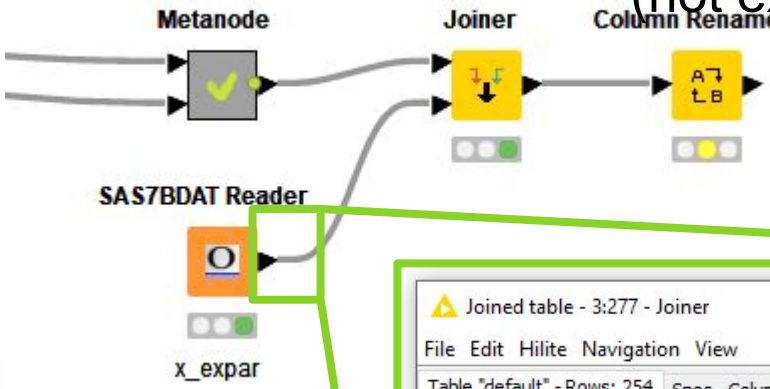
- Immediate clarity of algorithmic approach for anyone with programming / data science backgrounds
- Optimization of workflows straight forward

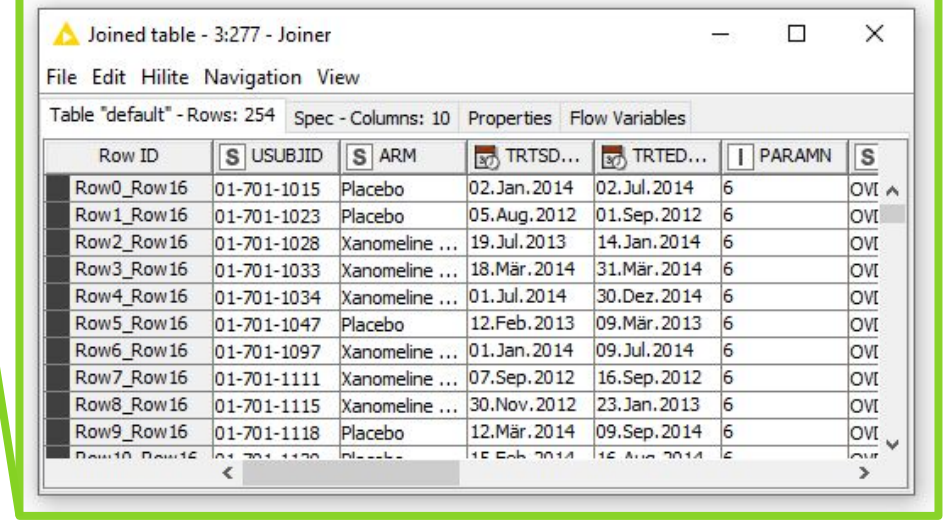
SAS vs. KNIME

Meta node (comparable to SAS Macros)

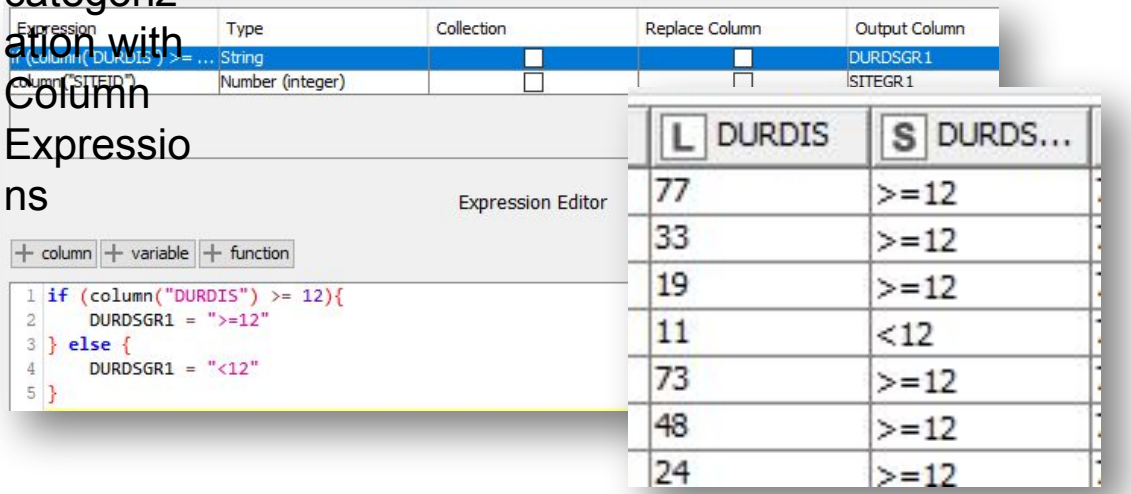
Configured node (not executed)

Data flows

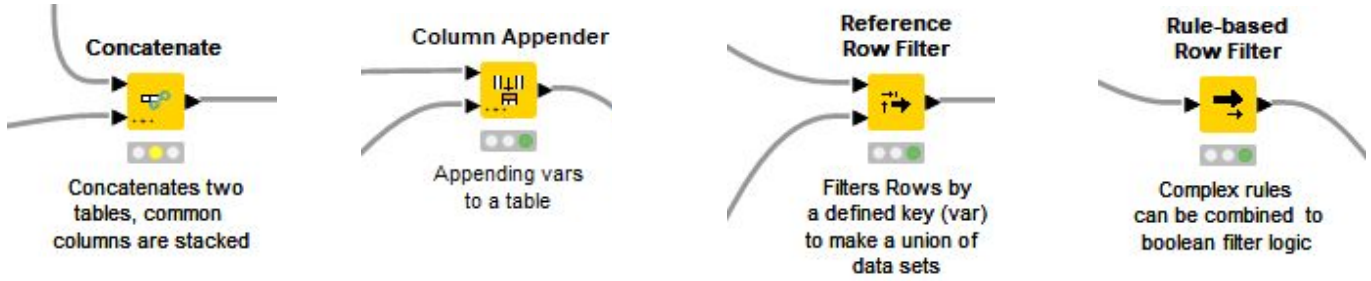




Example: Batch categorization with Column Expressions



Commonly required ETL functionality

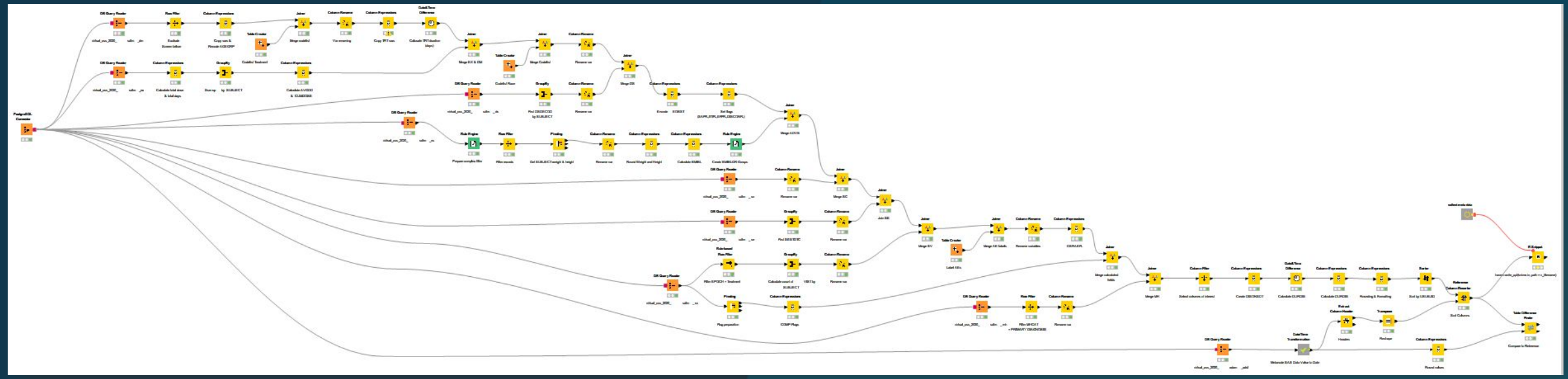


ADaM Derivation – ADSL *(subject listings)*

Data retrieval

Derivation

Testing
& XPT output

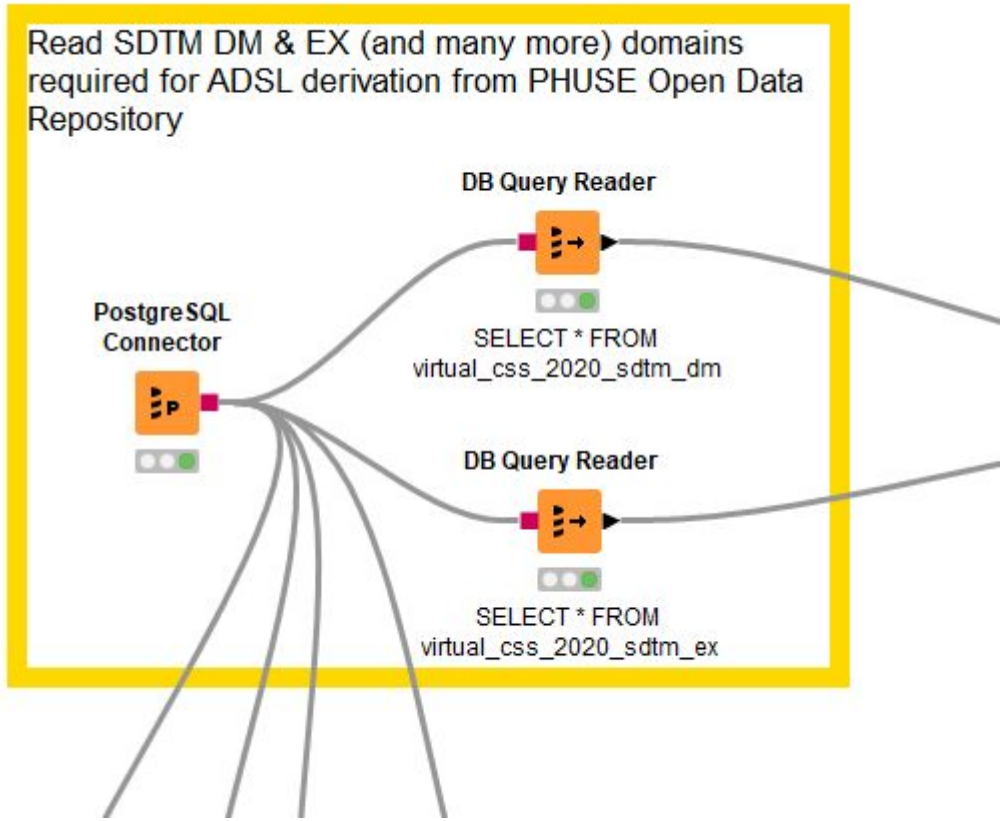


ADSL – Data Retrieval

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

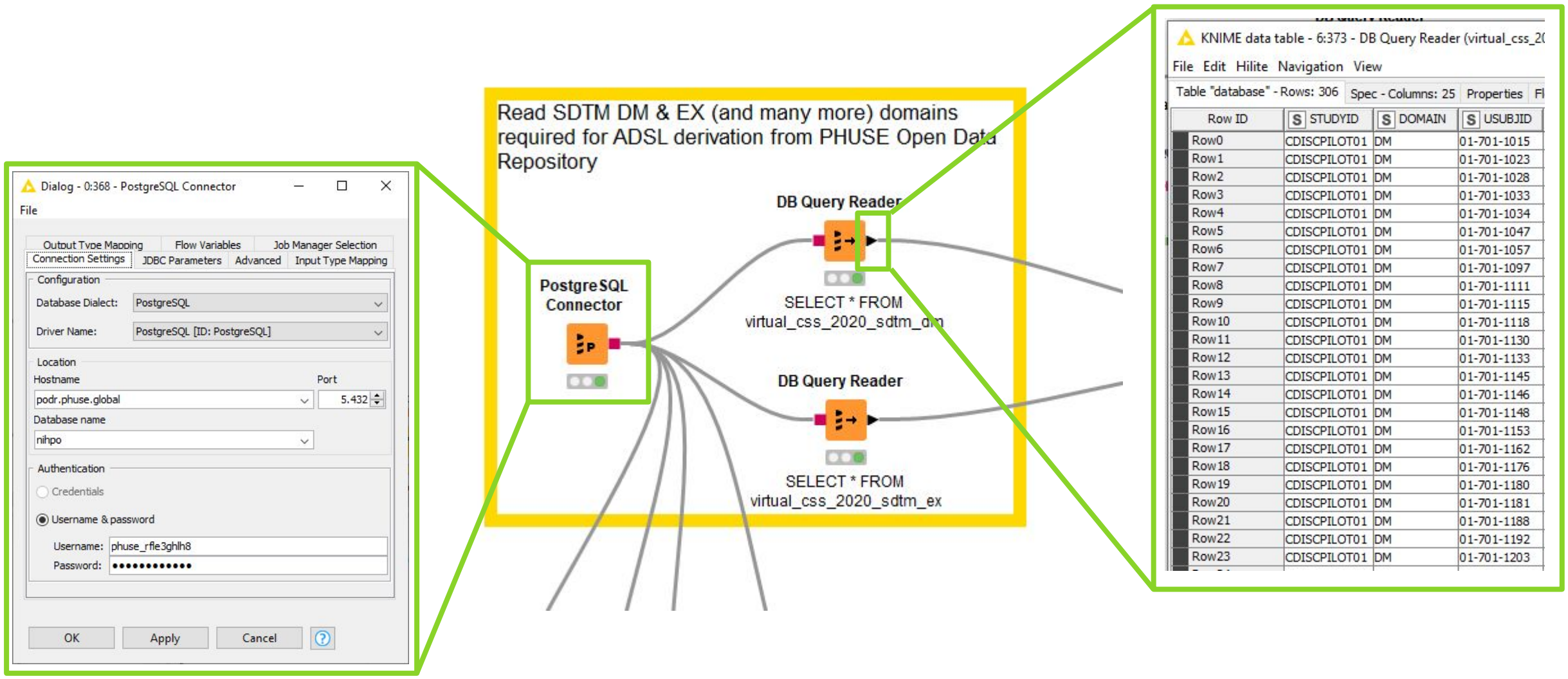
phuse | Open Data Repository

- *"PODR integrates health-related Open Data across agencies"*
- Sample data sets
  - SDTM, ADaM … and more



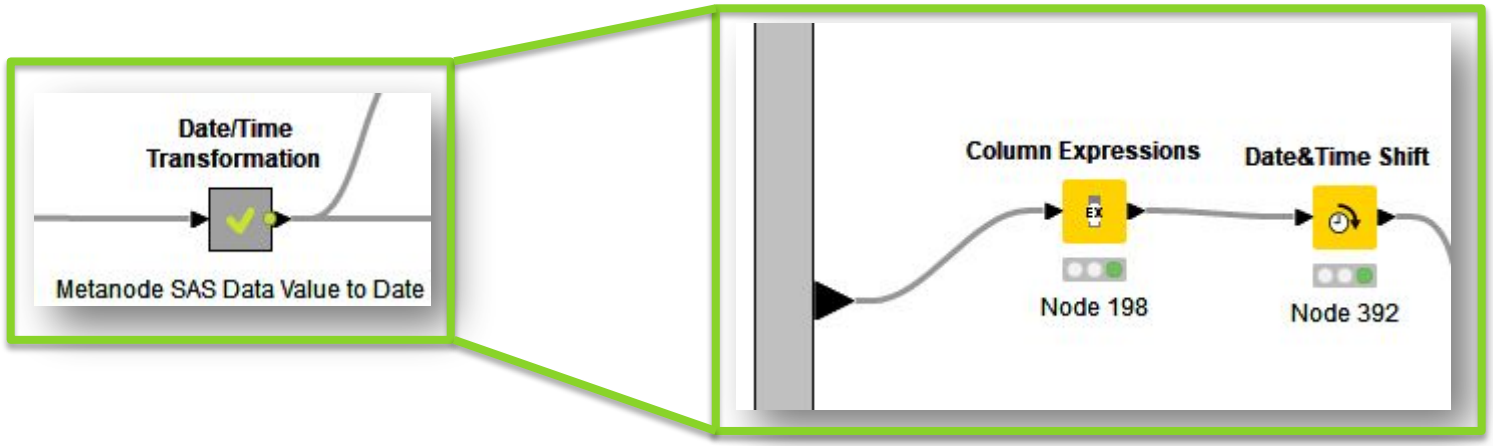
ADSL – Data Retrieval

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



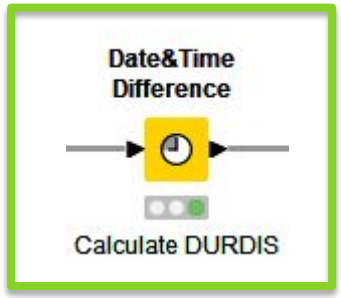
ADSL – Derivation

Coping with date formats



Allow date calculations

…



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

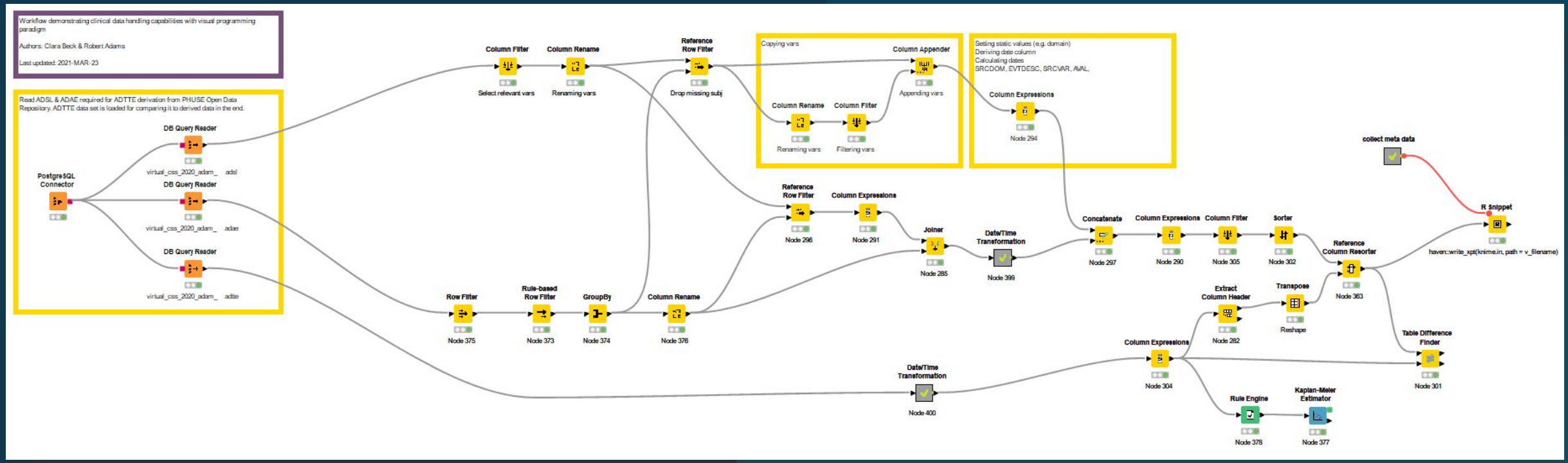
*TRTSTD = Treatment Start Date

ADaM Derivation – ADTTE (time to event)

Data retrieval

Derivation

Visualization & Testing & XPT output



Breakout – Survival Analysis

$$S(t) = P(T \geq t),$$

$$\hat{S}(t) = \prod_{i=1}^{t_i \leq t} \left(\frac{n_i - d_i}{n_i} \right)$$

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 surivival
  - 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 **harzard 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 (H_0: groups have the same hazard)
  - Usually Kaplan-Meier plots to present survival data are used

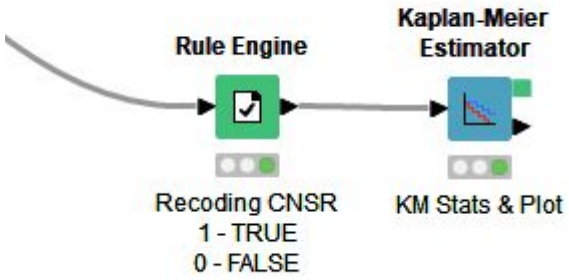
https://doi.org/10.1016/j.clon.2020.07.014

ADTTE – Derivation

Kaplan-Meier Statistics / Plotting

$$S(t) = P(T \geq t),$$

$$\hat{S}(t) = \prod_{i=1}^{t_i \leq t} \left(\frac{n_i - d_i}{n_i} \right)$$



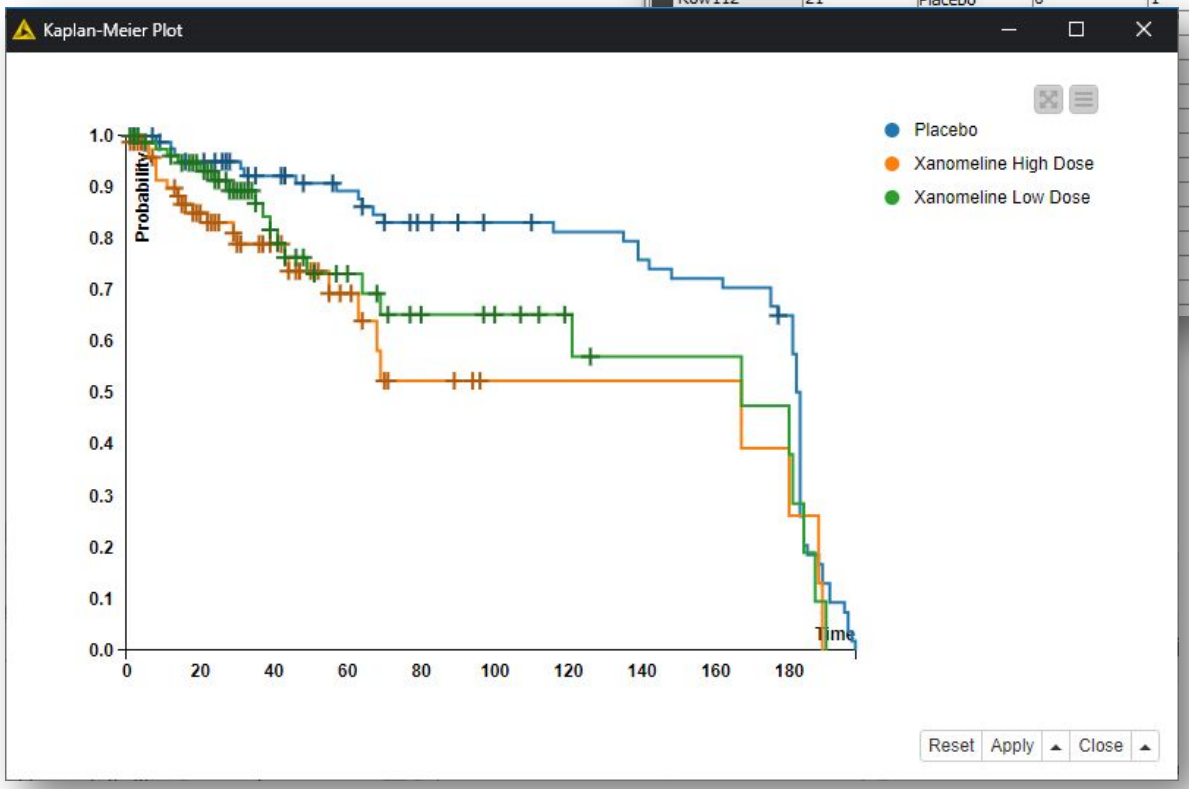
Caution: Details matter!
E.g. CI calculations have subtle changes in different programming languages and might have specific definitions

Statistics - 0:379 - Kaplan-Meier Estimator

File Edit Hilite Navigation View

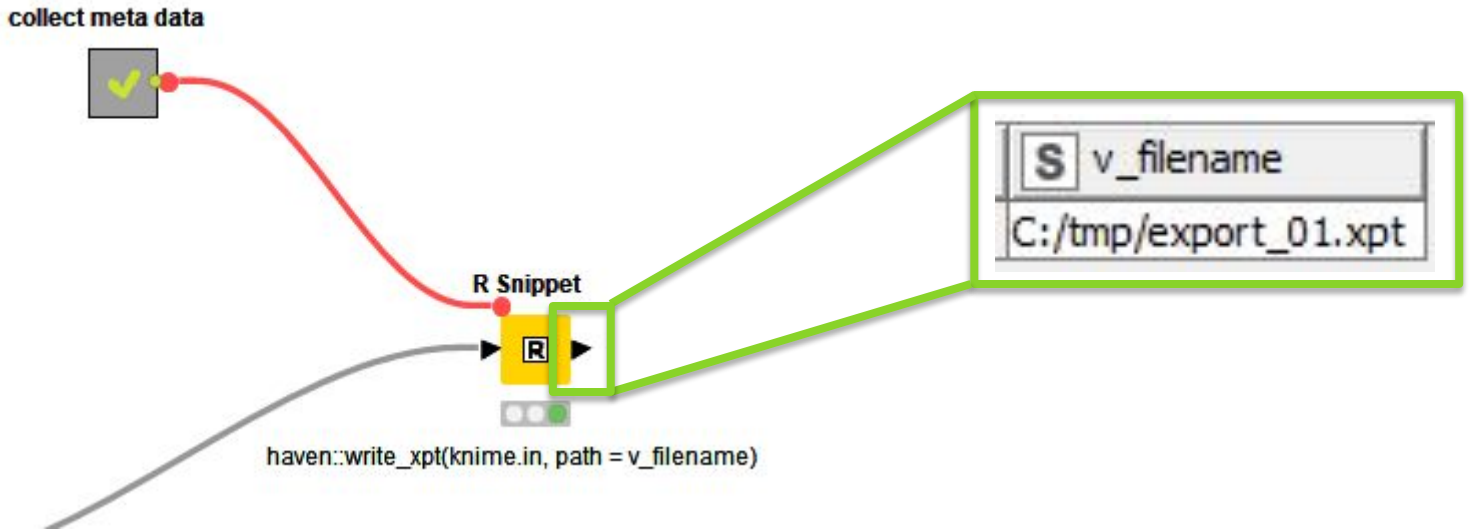
Table "default" - Rows: 161 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	AVAL	TRTA	#True(...	#False(...	km-esti...
Row152	1	Placebo	0	1	1
Row0	2	Placebo	0	1	1
Row1	3	Placebo	0	2	1
Row134	7	Placebo	0	1	1
Row68	8	Placebo	1	0	0.988
Row159	9	Placebo	0	1	0.988
Row39	12	Placebo	1	0	0.975
Row116	13	Placebo	1	0	0.963
Row29	14	Placebo	1	0	0.95
Row144	16	Placebo	0	1	0.95
Row112	21	Placebo	0	1	0.95
					0.95
					0.95
					0.95
					0.95
					0.936
					0.923
					0.923
					0.923
					0.923
					0.923
					0.908
					0.908

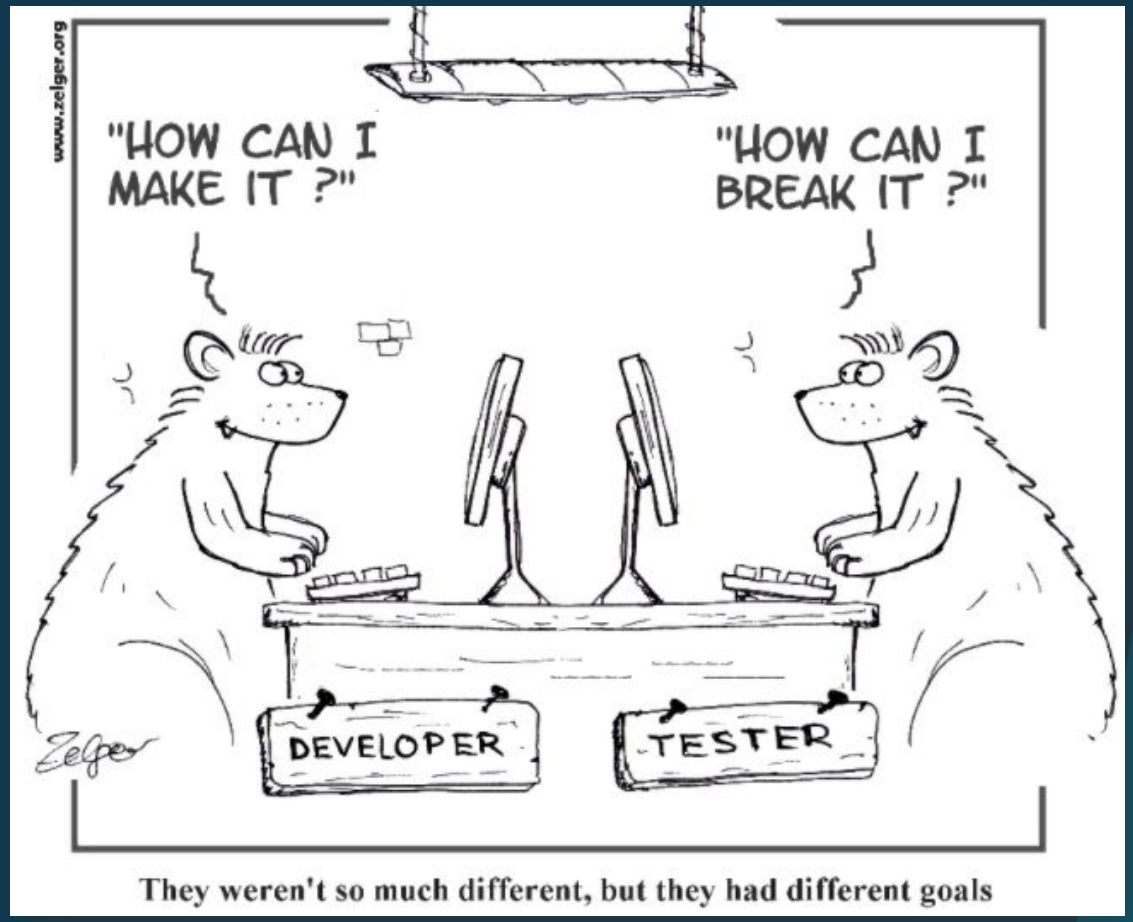


ADTTE – Derivation

Utilizing R to write XPT output (regulatory authority requirement)



Derived, tested table



Core of GxP:

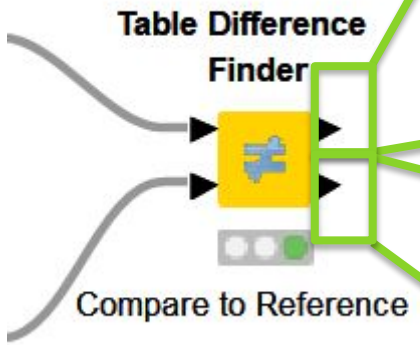
Validation,
Testing,
Logging,
Reporting

Double Programming Validation

Dear PODR – no offense ☺ you have a typo!

Derived data table

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



Value differences

RowId ...	RowId ...	Column...	Value Compared Table	Value Reference Table
Row0_Row0	Row0	EVNTDESC	Dermatologic Event Occu...	Dematologic Event Occu...
Row1_Row1	Row1	EVNTDESC	Dermatologic Event Occu...	Dematologic Event Occu...
Row2_Row2	Row2	EVNTDESC	Dermatologic Event Occu...	Dematologic Event Occu...
Row4_Row3	Row4	EVNTDESC	Dermatologic Event Occu...	Dematologic Event Occu...
Row6_Row4	Row6	EVNTDESC	Dermatologic Event Occu...	Dematologic Event Occu...
Row8_Row5	Row8	EVNTDESC	Dermatologic Event Occu...	Dematologic Event Occu...
Row10_Row6	Row10	EVNTDESC	Dermatologic Event Occu...	Dematologic Event Occu...
Row11_Row7	Row11	EVNTDESC	Dermatologic Event Occu...	Dematologic Event Occu...

Domain differences

PARAMCD	PARAMCD	true	true	true
AVAL	AVAL	true	false	true
STARTDT	STARTDT	true	true	true
CNSR	CNSR	true	true	true
EVNTDESC	EVNTDESC	true	false	true

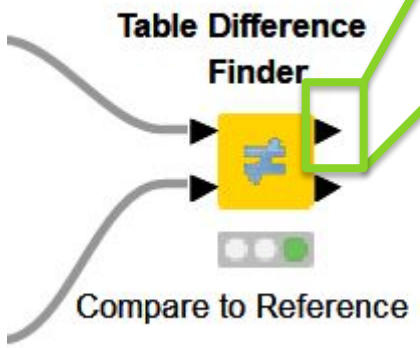
For numeric: min / max (range) differs
For nominal: different unique values

Domain differences are an additional validation layer

Double Programming Validation

Derived data table

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

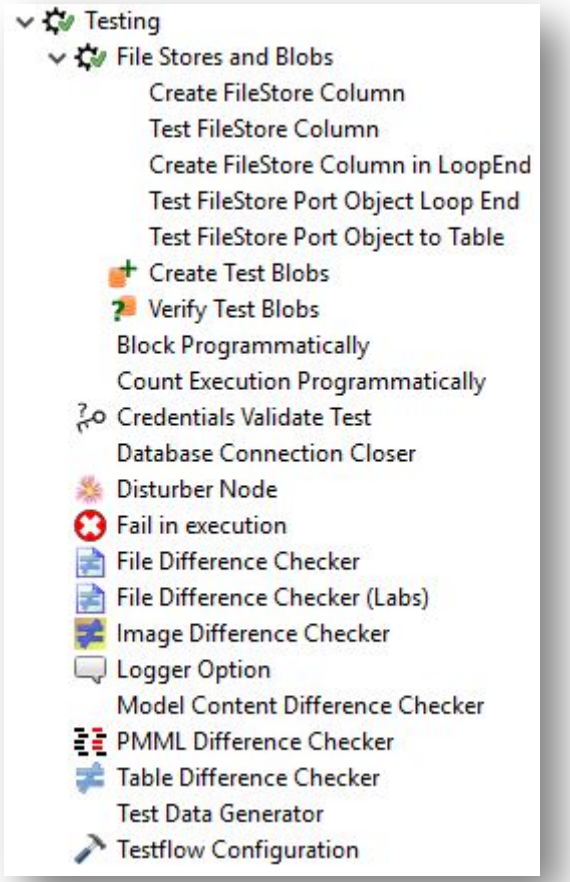


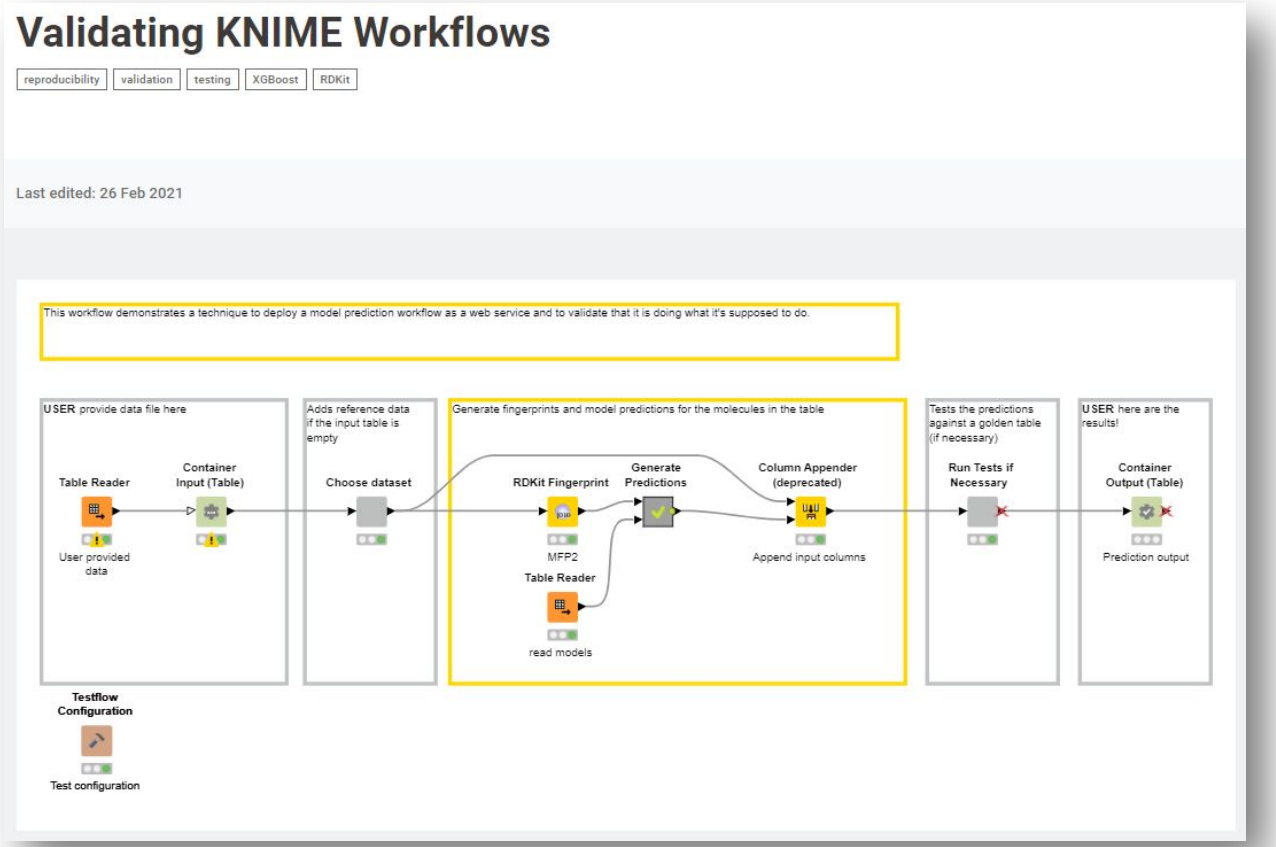
HEIGHTBL	135.89	135.9
WEIGHTBL	55.79	55.8
EDUCLVL	15	15.0
DURDIS	44	43.9

Value differences (e.g. rounding errors,
or different data types [int vs. double])

Testing

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



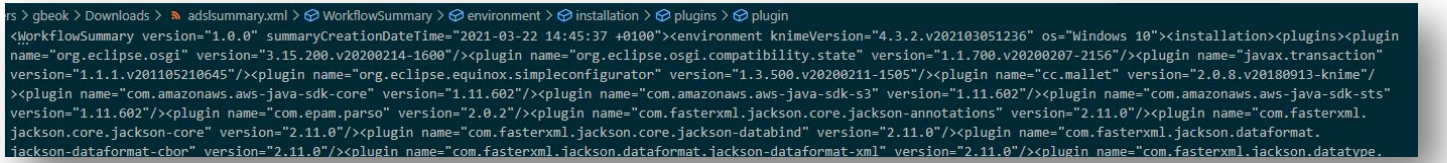


All relevant testing scenarios can be covered

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

org.knime.workbench.workflo...	4.3.0.v202011191609
org.knime.workflow.migration	4.3.0.v202011191336
org.knime.xml	4.3.0.v202011212018
org.mortbay.jetty.server	6.1.26
org.mortbay.jetty.util	6.1.26
org.mozilla.javascript	1.7.5.v201504281450
org.objectweb.asm	5.0.4
org.openqa.selenium	3.12.0.v202010300...
org.reactivestreams.reactive-...	1.0.3
org.sat4j.core	2.3.5.v201308161310

Installed plugins

Column Expressions	Manipulator	13
Column Filter	Manipulator	2
Column Rename	Manipulator	10
DB Query Reader	Source	9
Date&Time Differ...	Manipulator	2
Date/Time Shift (l...	Manipulator	4
Extract Column H...	Manipulator	1
Extract Context P...	Source	1
Extract System Pr...	Source	1
GroupBy	Manipulator	4
Java Snippet (sim...	Manipulator	1

Used nodes

Knime Version	OS	WF Author	Creatio...	Last Edi...	Last Up...	WF Description	summaryCreationDate...	WF Title	WF An...
4.3.2.v202103051236	Windows 10	Robert Adams & Clara Beck	2021-03-02	?	?	PHUSE_Workflow_ADSL_PODR Demonstrates ADSL derivation from PODR data sets for PHUSE US Connect 2021	2021-03-23 11:44:46 +0100	PHUSE_Workflow_ADSL_PODR	[]

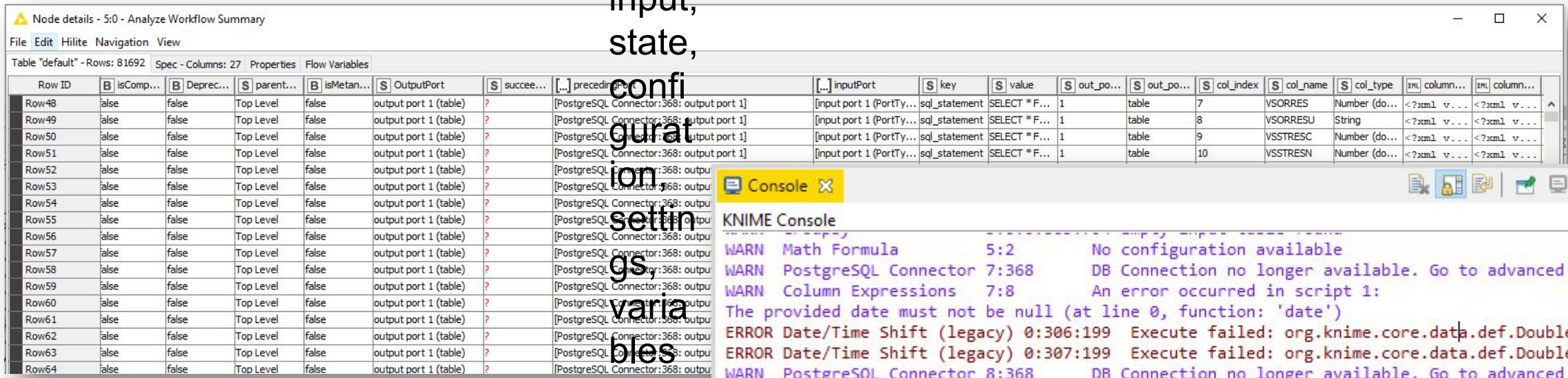
Environment & Metadata

Logging

Export executed workflow summary

Node details:

- For each node input, state, configuration, settings, variables
- Data at input



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

Clara Beck

Senior Computational Scientist

Bayer – Oncology Digitalization & Computational Science

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