DKTK German Cancer Consortium

Transforming EDC clinical data exports into reporting friendly data tables

Tomas Skripcak

German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), University Hospital Carl Gustav Carus, Technische Universität Dresden

LibreClinica Workshop 30.09.2022

INTRODUCTION TO DATA EXPORTS MODULE

LibreClinica relational database stores data for all hosted studies

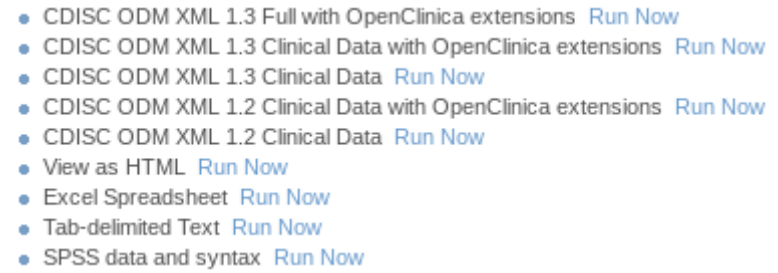
- data in long table “item_data” -> Entity-Attribute-Value (EAV)
- linked metadata tables (CDISC ODM XML standard)
- export generate ODM XML of specific type (full, clinical_data)
- ODM XML is enhanced with vendor specific extensions

- extract typically creates very wide table data representation
- data extract architecture allows to create new export formats without a need for application recompilation and packaging
- XSL stylesheet transformations: ODM XML -> specific output format
- optional execution of post processing (sql, pdf)

INTRODUCTION TO DATA EXPORTS MODULE

Default EDC -> 9 export formats

- libreclinica.config/extract.properties
- configuration is parsed on system start
- extracts are enumerated in config (1-9)
- use incremental numbers for new extract
- important “odmType” and “file” attributes
- multiple .xsl -> multiple output files for one export format (e.g. SPSS)
- postprocessors are disabled (bug free?)
- new postprocessors require code changes



Folder: libreclinica.data/xslt

- copyXML.xsl
- copyXML.xsl
- odm1.3_to_1.3_no_extensions.xsl
- odm1.3_to_1.2_extensions.xsl
- odm1.3_to_1.2.xsl
- odm_to_html.xsl
- ODMToTAB.xsl
- ODMToTAB.xsl
- odm_spss_sps.xsl, odm_spss_dat.xsl

DATA EXPORTS PROS AND CONS

Tab delimited text and Excel formats are versatile tabularisation options

- wide table data representation (one row per subject)
- guarantee unique column names
  - ItemName_Ea(_b)_Cc(_d)
    - a = event number
    - b = event repeat key (for repeating events only)
    - c = CRF number
    - d = item group repeat key (for repeating groups aka grids)
- mid study protocol changes -> a, c numbers of specific items in export can change -> analysis scripts need adaptations
- good to know: dates are ISO 8601 formatted, decimals use “.” separator, multi value “,” separator

DATA EXPORTS PROS AND CONS

- Excel with German locale -> auto formatting issues
- TSV is easier to deal with (LibreOffice, R or Python)
- number of columns in dataset varies depending on completeness of data entry
- variables with no value across whole cohort are not exported
- the only way of getting complete dataset is to ensure that all show/hide logic question combinations and all optional fields have value for at least one subject
- introduction of dummy subjects that need to be removed in post processing
- multiple eCRF versions in export require post processing

DATA MARTS

Concept of data mart is based on idea of splitting the one database scheme of EDC system into separate scheme per each study, where data is represented in consistent way with stable attribute naming to allow easier queries, reporting and post processing tasks

- created by ETL process as secondary data store
- snapshot of data (not always in sync with EDC system)
- allow multiple views on data (long vs wide table)
- allow complex long running queries without stressing EDC system
- present read only data to new type of users (not restricted by strict EDC user roles)

DATA MARTS

Each eCRF version/ item group is represented in its own table (pivot), respecting the data type for each item, but prevents very wide tables

eCRF/Event	6-W-FUP	FUP
FUP-SV	x	x

SSID	Visit	RepeatKey	Form	NY_VISIT	VISITDAT	VISITREASND	VISIT_TP	TPU
ABX001	6-W-FUP	1	FUP-SV	1	2016-11-04		6	W
ABX002	FUP	1	FUP-SV	1	2016-11-25		3	M
ABX002	FUP	2	FUP-SV	1	2017-01-16		6	M

* additional tables can provide list of enrolled subjects, events, form statuses and subject groups/arms

LESSON LEARNED – WHY CUSTOM DATA MART

Not everybody needs a custom build data mart solution -> check the existing community projects for OpenClinica/LibreClinica.

Reasons why we implemented one:

- solution not bound to SQL database of one specific EDC system only
- driven by ODM XML standard to stay vendor agnostic
- freedom in choice of storage and presentation layer
- close integration into existing IT infrastructure
- choice of post processing scripting tools (SQL vs R vs Python)

LESSON LEARNED – STORAGE AND PRESENTATION

Backend storage is traditionally relational however there are multiple systems that can be utilised for this purpose

- pure storage oriented: Access, PostgreSQL, …
- storage with presentation layer: Metabase, LabKey, …

Vendor agnostic base ETL with ability to include vendor specific details

- base: parsing ODM XML exports (from any system)
- extended: recognizing vendor specific extensions in ODM XML file
- integrated: querying system database for extras that are not in ODM XML

LESSON LEARNED - CONVENTIONS

- eCRF table names and item names correspond to ODM definition
- descriptions used to enhance metadata representation (tooltips)
- OIDs uses internally as identifier and as foreign key relationships
- use metadata to present all items in table (also items without values)
- 2 views on repeating groups
  - wide table – part of eCRF table
  - long table – repeating item group on its own
- include both coded as well as decoded item values
- multi selects transformed into Boolean columns
- partial dates represented as full dates with min/max range
- usefull attribute tables for subjects, events, forms and subject groups with statuses

EXAMPLES – LABKEY TABLES AND REPORTS

	All Visits	Enrollment	Baseline	Treatment	End-of-Therapy	TEL-FUP	QLQ-FUP	6-W-FUP	Follow-Up	End-of-Study	Adverse-Event	Death-Details	Drop-Out
EDC-Attributes													
SubjectAttributes	45	45											
SubjectGroupAttributes	27	27											
EventAttributes	274		43	24	29	27	29	30	28	23	24	2	15
FormAttributes	270		43	24	29	27	29	28	27	23	23	2	15
EDC-FormVersions													
DM - DE v.1.0	42		42										
PX-REG - DE v.1.0	43		43										
TX-WEEK - DE v.1.0	49		25	24									
ECOG - DE v.1.0	131		26	24	26			28	27				
PX-ATOX - DE v.1.0	78		23	24	6			25					
ATOX - DE v.1.0	12		5	1	2			4					
QLQ-C30 - DE v.1.0	90		34		27		29						
QLQ-LC13 - DE v.1.0	90		34		27		29						
VS - DE v.1.0	26				26								
PX-RTX - DE v.1.0	25				25								
FUP-SV - DE v.1.1	111					27	29	28	27				
PX-TEL - DE v.1.0	27					27							
MORTALITY - DE v.1.0	55							28	27				
PX-CLS - DE v.1.0	55							28	27				
PX-TRS - DE v.1.0	55							28	27				
LTOX-THO - DE v.1.0	27								27				
LTOX - DE v.1.0	6								6				
EOS - DE v.1.0	23									23			
AE - DE v.1.0	21										21		
SAE - DE v.1.0	21										21		
AE-EVL - DE v.1.0	23										23		
DEATH - DE v.1.0	2											2	
DROP-OUT - DE v.1.0	15												15
EDC-ItemGroups													
GICDCOD - ICD Causes of death - [IG_DEATH_GICDCOD]	1											1	
GATOX - Acute Toxicities - [IG_ATOX_GATOX]	5			1	2			2					
GLTOX - Late Toxicities - [IG_LTOX_GLTOX]	2								2				
GRTXINT - Bestrahlung unterbrochen - [IG_PXRTX_GRTXINT]	5				5								
GSCTX - Simultane Chemotherapie - [IG_PXRTX_GSCTX]	25				25								
GACTX - Adjuvante Chemotherapie - [IG_PXRTX_GACTX]	4				4								
GCTX - Chemotherapie - [IG_PXTRS_GCTX]	11								11				

EXAMPLES – LABKEY TABLES AND REPORTS

Dataset: QLQ-C30 - DE v.1.0, All Visits

Version Description: EORTC QLQ-C30 (version 3.0) Revision Notes: First Version (DE)

Manage Groups View Specimens

SSID	Visit	StudyEventOID	StudyEventRepeatKey	StudySiteIdentifier	FormOID	PHYS_DEMANDING	LONG_WALK	SHORT_DISTANCE	BED	HELP	LIMITED_WORK	LIMITED_HOBBY	BREATHLESS	PAIN	REST	SLEEPING_DISORDER	WEAK	APPETITE	SICK	VOMIT
	Baseline	SE_PXBASELINE	1		F_QLQC30_DEV10	2	2	1	1	1	2	2	2	1	2	3	1	2	1	2
	End-of-Therapy	SE_PXEOT	1		F_QLQC30_DEV10	4	4	1	3	1	2	4	3	3	4	1	3	4	2	1
	QLQ-FUP	SE_PXQLQFUP	1		F_QLQC30_DEV10	3	2	1	2	1	2	2	2	1	2	2	2	3	1	1
	QLQ-FUP	SE_PXQLQFUP	4		F_QLQC30_DEV10	2	1	1	1	1	2	2	4	3	2	2	2	1	1	1
	QLQ-FUP	SE_PXQLQFUP	5		F_QLQC30_DEV10	2	1	1	2	1	2	2	2	3	2	2	2	1	1	1
	QLQ-FUP	SE_PXQLQFUP	11		F_QLQC30_DEV10	3	1	2	2	1	2	2	2	1	2	3	2	1	2	1
	Baseline	SE_PXBASELINE	1		F_QLQC30_DEV10	2	3	1	1	1	1	2	2	1	1	2	2	1	1	1
	End-of-Therapy	SE_PXEOT	1		F_QLQC30_DEV10	2	2	2	2	1	2	2	4	4	3	4	3	2	1	1
	QLQ-FUP	SE_PXQLQFUP	1		F_QLQC30_DEV10	2	3	1	3	1	3	2	3	1	3	4	3	2	2	1
	QLQ-FUP	SE_PXQLQFUP	2		F_QLQC30_DEV10	2	2	1	2	1	2	2	2	2	2	2	2	1	1	1
	QLQ-FUP	SE_PXQLQFUP	3		F_QLQC30_DEV10	2	1	1	1	1	1	1	2	3	1	3	1	1	1	1
	QLQ-FUP	SE_PXQLQFUP	4		F_QLQC30_DEV10	3	2	1	1	1	2	2	3	3	2	3	2	1	1	1
	Baseline	SE_PXBASELINE	1		F_QLQC30_DEV10	2	2	1	1	1	1	1	2	2	1	2	2	1	1	1
	End-of-Therapy	SE_PXEOT	1		F_QLQC30_DEV10	1	1	1	1	1	1	1	1	1	2	2	2	1	1	1
	QLQ-FUP	SE_PXQLQFUP	1		F_QLQC30_DEV10	2	2	1	1	1	2	2	2	1	2	1	2	1	1	1
	QLQ-FUP	SE_PXQLQFUP	2		F_QLQC30_DEV10	2	1	1	1	1	1	1	1	1	2	1	1	1	1	1
	QLQ-FUP	SE_PXQLQFUP	3		F_QLQC30_DEV10	2	1	1	1	1	1	1	1	1	2	1	1	1	1	1

EXAMPLES – LABKEY TABLES AND REPORTS

Dataset: EventAttributes, All Visits

Contains up to one row of Participants data for each Participant/Visit combination.

Manage Groups View Specimens 1 - 100 of 894

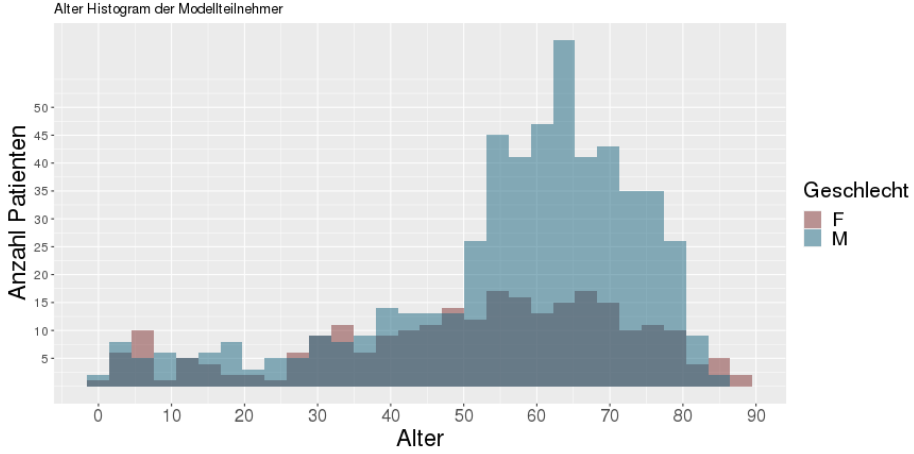
SSID	Visit	StudyEventOID	EventName	StartDate	EndDate	Status	SystemStatus	StudyEventRepeatKey	Type
	Baseline	SE_PXBASELINE	Baseline	2016-08-03 00:00		completed	available	1	scheduled
	Treatment	SE_PXTRE	Treatment	2016-08-26 00:00		completed	available	1	scheduled
	Treatment	SE_PXTRE	Treatment	2016-09-01 00:00		completed	available	2	scheduled
	Treatment	SE_PXTRE	Treatment	2016-09-08 00:00		completed	available	3	scheduled
	Treatment	SE_PXTRE	Treatment	2016-09-15 00:00		completed	available	4	scheduled
	Treatment	SE_PXTRE	Treatment	2016-09-22 00:00		completed	available	5	scheduled
	Treatment	SE_PXTRE	Treatment	2016-09-29 00:00		skipped	unavailable	6	scheduled
	End-of-Therapy	SE_PXEOT	End-of-Therapy	2016-09-23 00:00	2016-09-23 00:00	completed	available	1	scheduled
	QLQ-FUP	SE_PXQLQFUP	QLQ-FUP	2016-11-04 00:00		completed	available	1	scheduled
	QLQ-FUP	SE_PXQLQFUP	QLQ-FUP	2016-11-25 00:00		skipped	unavailable	2	scheduled
	QLQ-FUP	SE_PXQLQFUP	QLQ-FUP	2017-03-23 00:00		skipped	unavailable	3	scheduled
	QLQ-FUP	SE_PXQLQFUP	QLQ-FUP	2017-06-06 00:00		completed	available	4	scheduled
	QLQ-FUP	SE_PXQLQFUP	QLQ-FUP	2017-09-13 00:00		completed	available	5	scheduled
	QLQ-FUP	SE_PXQLQFUP	QLQ-FUP	2017-12-15 00:00		skipped	unavailable	6	scheduled
	QLQ-FUP	SE_PXQLQFUP	QLQ-FUP	2018-03-15 00:00		skipped	unavailable	7	scheduled
	QLQ-FUP	SE_PXQLQFUP	QLQ-FUP	2018-06-15 00:00		skipped	unavailable	8	scheduled
	QLQ-FUP	SE_PXQLQFUP	QLQ-FUP	2018-09-14 00:00		skipped	unavailable	9	scheduled
	QLQ-FUP	SE_PXQLQFUP	QLQ-FUP	2019-10-04 00:00		skipped	unavailable	10	scheduled
	QLQ-FUP	SE_PXQLQFUP	QLQ-FUP	2020-09-30 00:00		completed	available	11	scheduled
	QLQ-FUP	SE_PXQLQFUP	QLQ-FUP	2022-01-04 00:00		scheduled	available	12	scheduled
	6-W-FUP	SE_PX6WFUP	6-W-FUP	2016-11-04 00:00		completed	available	1	scheduled
	Follow-Up	SE_PXFUP	Follow-Up	2016-11-25 00:00		completed	available	1	scheduled
	Follow-Up	SE_PXFUP	Follow-Up	2017-03-23 00:00		skipped	unavailable	2	scheduled
	Follow-Up	SE_PXFUP	Follow-Up	2017-06-06 00:00		completed	available	3	scheduled
	Follow-Up	SE_PXFUP	Follow-Up	2017-09-13 00:00		completed	available	4	scheduled
	Follow-Up	SE_PXFUP	Follow-Up	2017-12-15 00:00		skipped	unavailable	5	scheduled

EXAMPLES – LABKEY TABLES AND REPORTS

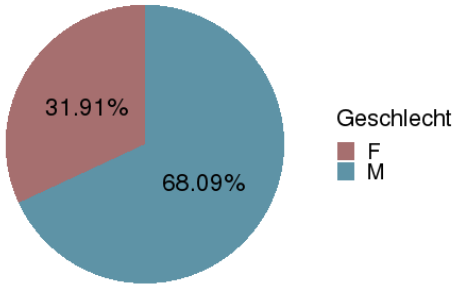
Soziodemographische Merkmale

Etwa zwei Drittel (68.09 %) der 799 Modellteilnehmer sind Männer, etwa ein Drittel (31.91 %) sind Frauen. Das Durchschnittsalter lag bei 54.94 Jahren,

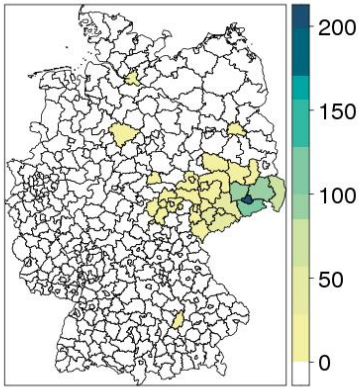
Alter Histogram der Modellteilnehmer



Geschlecht der Modellteilnehmer



Geographishe Verteilung der Modellteilnehmer



region	value
Dresden	199
Landkreis Sächsische Schweiz-Osterzgebirge	117
Landkreis Bautzen	112
Landkreis Meißen	91
Landkreis Görlitz	68
Landkreis Mittelsachsen	53
Erzgebirgskreis	32
Landkreis Zwickau	29
Leipzig	15
Landkreis Leipzig	12
Vogtlandkreis	12
Landkreis Nordsachsen	11
Chemnitz	10

THANKS FOR YOUR ATTENTION

- Developed by Dresden IT team in scope of DKTK RadPlanBio project:
  - Ronny Kursawe
  - Tomas Skripcak

- Study data examples are courtesy of OncoRay Clinical Trial Centre operated under umbrella of:
  - the translation centre of DKTK partner site Dresden established jointly by the University Hospital Carl Gustav Carus at the Technische Universität Dresden, Institution under public law of the free state of Saxony