



Generating Synthetic eCOA Non-CRF Data from External Vendors Using Python

JASKARAN SINGH SAINI
GENINVO, CHANDIGARH, INDIA



PP18

Introduction:

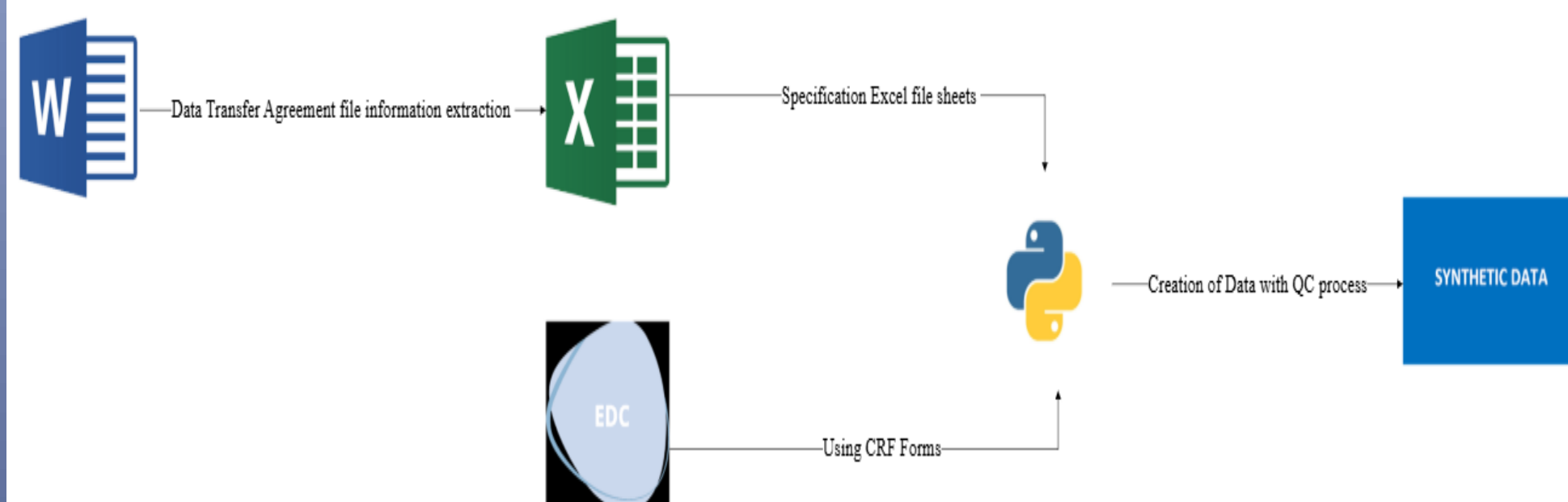
In clinical research, the integration of Non-Case Report Form (Non-CRF) data from external vendors, such as Electronic Clinical Outcome Assessment (eCOA) data, presents significant challenges. These challenges include ensuring data privacy, maintaining data integrity, handling missing data, non-standardization in Data Transfer Agreement (DTA) and Data Transfer Specification (DTS) files from vendors, and achieving interoperability with existing clinical trial data.

This poster demonstrates the creation of synthetic ePRO non-CRF data from an eCOA collecting system, specifically in the QS (Questionnaires domain) form, using open-source technology Python and its various packages (Pandas, NumPy, etc.). The process also incorporates EDC (Electronic Data Capture) eCRF data such as DM (Demographic Data) and SV (Schedule of Visit) forms, and in some cases, uses leading question variables from respective forms. In the data generation process, we create a standard Excel file from DTA (Data Transfer Agreement) files and utilize all the gathered information to generate meaningful synthetic data that mirrors real-world data from external vendors. Using this process, the creation of synthetic data for other domains like FA, RS, and LB is feasible with multiple use cases.

DESIGN:

The QS (Questionnaire) in CDISC standards is a findings domain form, that contains data for named, stand-alone instruments designed to provide an assessment of a concepts. In laymen terms it is designed to capture and manage data collected from questionnaires used in clinical trials. In this one record per questionnaire per question per time point per visit per subject is captured. These structured instruments assess various concepts, such as symptoms, behaviors, or quality of life, through a series of related questions. When integrated with electronic Clinical Outcome Assessments (eCOA), which collect patient-centered outcomes electronically, the QS domain enhances the efficiency, accuracy, and comprehensiveness of data collection.

Process Flow:



The synthetic data creation process for QS form is divided in the following steps :

- 1.Data Transfer Agreement file information extraction
- 2.Creation of a Specification (excel file) having various sheets like Metadata, Timepoints etc.
- 3.Data files required from CRF (EDC) data SV, DM, QS (if present) file for visit, demographic and leading question information.

In Specification (excel file), the created sheets are listed below with examples:

Metadata:									
QS- Questionnaire									
Variable Order	Variable Name	Variable Label	Format	Anticipated Maximum Length	Description of Variable	Core	Example Values	Tests	Codelists
1	STUDYID	Study Identifier	Char	10	Unique identifier for a study.	Req	CDISC01		
2	DOMAIN	Domain Abbreviation	Char	2	Two-character abbreviation for the domain.	Req	QS		
3	USUBJID	Unique Subject Identifier	Char	20	Identifier used to uniquely identify a subject.	Req	CDISC01-1001001		
4	QSGRPID	Group ID	Char	20	Used to tie together a block of related records in a single domain for a	Perm			
5	QSTESTCD	Question Short Name	Char	8	Short name for the value in QSTEST, which can be used as a column h	Req			
6	QSTEST	Question Name	Char	40	Verbatim name of the question or group of questions used to obtain th	Req		Test	
7	QSCAT	Category of Question	Char	200	Used to specify the questionnaire in which the question identified by Q	Req		Test	
8	QSSCAT	Subcategory for Question	Char	200	A further categorization of the questions within the category. Example	Perm		Test	
9	QSORRES	Finding in Original Units	Char	200	Finding as originally received or collected (e.g., "RARELY", "SOMETIME	Exp		Test	
10	QSORRESU	Original Units	Char	50	Original units in which the data were collected. If not collected in orig	Perm			Code
11	QSSTRES	Character Result/Finding	Char	200	Contains the finding for all questions if responses copied or derived fr	Perm		Test	
12	QSSTRESN	Numeric Finding in Stand	Num	8	Used for continuous or numeric findings if standard format copied in	Perm		Test	
13	QSSSTAT	Completion Status	Char	20	Indicates whether a question was answered or not answered.	Perm	NOT DONE		
14	QSSREASON	Reason Not Answered	Char	200	Describe why a question was not answered. Used in conjunction with	Perm	Subject Refused		

Visit:				
VISIT				
ORDER	VISIT	VISITNUM	QSTPT	QSTPTNUM
1	SCREEN	2000		
2	DAY1	2001	PRE	1
3	DAY1	2001	POST	2
4	WEEK2	2002		
5	WEEK3	2003		
6	WEEK4	2004		
7	WEEK5	2005		
8	WEEK6	2006		
9	WEEK7	2007		
10	WEEK8	2008		
11	WEEK9	2009		
12	WEEK10	2010		

Test:							
TEST							
QSTESTCD	QSTEST	QSCAT	QSSCAT	QSORRES	QSORRESU	QSSTRES	QSSTRESN
SWLS0101	SWLS01-My Life is Close to Ideal	SWLS		Disagree		3	3
SWLS0101	SWLS01-My Life is Close to Ideal	SWLS		Neither Agree Nor Disagree		2	2
SWLS0101	SWLS01-My Life is Close to Ideal	SWLS		Agree		1	1
SWLS0102	SWLS02-Have Gotten Important Things	SWLS		Disagree		5	5
SWLS0102	SWLS02-Have Gotten Important Things	SWLS		Slightly Disagree		4	4
SWLS0102	SWLS02-Have Gotten Important Things	SWLS		Neither Agree Nor Disagree		3	3
SWLS0102	SWLS02-Have Gotten Important Things	SWLS		Slightly Agree		2	2
SWLS0102	SWLS02-Have Gotten Important Things	SWLS		Agree		1	1
SWLS0103	SWLS03-If I Could Live My Life Over	SWLS		Strongly disagree		6	6

Codelist :			
CODE LIST			
ORDER	CODELIST REFERENCE	VALUE	TEXT
1	APYN101	Y	YES
2	APYN101	N	NO
3	APFN102	None	NONE
4	APFN102	Weak	WEAK
5	APFN102	Moderate	MODERATE
6	APFN102	Strong	STRONG
7	APRT103	1	Very Difficult
8	APRT103	2	Difficult
9	APRT103	3	Moderate
10	APRT103	4	Easy
11	APRT103	5	Very Easy

4. Python algorithm processing and the steps for data creation:

- Metadata mapping
- Visit mapping
- CRF data processing
- DTA Automation process
- Data creation

5. Validation and QC process

The Python algorithm process the specification file sheets and required EDC data forms and caters the different mapping steps, and finally the dta automation process runs to create synthetic data.

```
import numpy as np
import pandas as pd
import os
import re
import datetime
from datetime import date, timedelta
import datetime as dt
from DTA function import *

#File Fetching Information
edc_data_sheet = edc_data_sheet(edc_location)
Spec_File = file_generation(File_location)
sheet_names = spec_sheet_name_extraction(Spec_File)

#Sheet Information Filtration
Metadata_Sheet = sheet_filtration(Spec_File, sheet_names[0])
Visit_Sheet = sheet_filtration(Spec_File, sheet_names[1])
Test_Sheet = sheet_filtration(Spec_File, sheet_names[2])
Codelist_Sheet = sheet_filtration(Spec_File, sheet_names[3])

#Data Processing
Metadata_sheet_mapping = metadata_mapping(Metadata_Sheet)
Visit_sheet_mapping = visit_mapping(Visit_Sheet, edc_data_sheet)
CRF_data_processing = crf_mapping(edc_data_sheet, Visit_sheet_mapping, Test_Sheet, "unique")

#Data generation
data= dta_auto_process(Metadata_sheet_mapping, Visit_sheet_mapping, CRF_data_processing)
save_dataframe_to_csv(data, location)
```

RESULT:

The QS synthetic data file:

	A	B	C	D	E	F	G	H	I	J	K
	STUDYID	DOMAIN	USUBJID	QSGRPID	QSTESTCD	QSTEST	QSCAT	QSSCAT	QSORRES	QSORRESU	QSSTRES
1	CDISC01	QS	CDISC01-1001001		SWLS0101	SWLS01-My Life is Close to Ideal	SWLS		Slightly agree		2
2	CDISC01	QS	CDISC01-1001002		SWLS0102	SWLS02-Have Gotten Important Things	SWLS		Neither agree nor disagree		4
3	CDISC01	QS	CDISC01-1001003		SWLS0103	SWLS03-If I Could Live My Life Over	SWLS		Agree		1
4	CDISC01	QS	CDISC01-1001004		SWLS0104	SWLS04-My Life Conditions are Excellent	SWLS		Disagree		5
5	CDISC01	QS	CDISC01-1001005		SWLS0105	SWLS05-Were you Tired	SWLS		Strongly disagree		6
6	CDISC01	QS	CDISC01-1001006		SWLS0106	SWLS06-Short of Breath in usual activities	SWLS		Slightly agree		1
7	CDISC01	QS	CDISC01-1001007		SWLS0107	SWLS07-Always feels low in energy	SWLS		Neither agree nor disagree		3
8	CDISC01	QS	CDISC01-1001008		SWLS0108	SWLS08-Do you feel depressed	SWLS		Agree		1
9	CDISC01	QS	CDISC01-1001009		SWLS0109	SWLS09-Overall Health	SWLS		1		1
10	CDISC01	QS	CDISC01-1001010		SWLS0110	SWLS10-Motivation in Everyday Life	SWLS		2		2
11	CDISC01	QS	CDISC01-1001011	APYN1101	APYN01-Do you have diabetes	LQLS			Y		Y
12	CDISC01	QS	CDISC01-1001012	APYN1102	APYN02-Do you have asthma	LQLS			N		N

Using the Validation and QC process QS form data is refined for the any further required usage. The usage of this synthetic external vendor data:

- Data Privacy and Security
- Scenario Simulation
- Quality Control and Testing
- Algorithm Training and Validation
- Improving Data Privacy
- Enhancing Data Utility

Using this process with further enhancement we can create Non-CRF data for FA, RS, LB etc domains also.

REFERENCES:

- [Use of external data to accelerate evidence generation - digital twins and external controls_H.Frohlich](#)
- [What is eCOA? | Clario](#)
- <https://www.certara.com/blog/all-you-need-to-know-about-non-crf-data-in-clinical-trials/>
- [Synthetic data as external control arms in scarce single-arm clinical trials | PLOS Digital Health](#)
- [The use of synthetic or external data in clinical trials | Blog](#)
- <https://clario.com/resources/articles/what-is-ecoa-and-how-does-it-improve-clinical-trial-data-quality/>

CONTACT:

LET'S CONNECT ON LINKDIN:

