



Leveraging Python for Synthetic External Vendor Data Generation for eCOA and ePRO

Jaskaran Singh Saini
Hitesh Raval

www.geninvo.com
North America | UK | Europe | India



Agenda

- 1 Background: eCOA/ePRO and External Data
- 2 Traditional Synthetic Data Approaches: Critical Limitations
- 3 Overcoming Barriers: A Python-Based Approach
- 4 Workflow: Steps for Synthetic Data Creation
- 5 Open-Source Python Stack
- 6 Python Algorithm Processing:
- 7 Results
- 8 Applications



www.geninvo.com
North America | UK | Europe | India



Background: eCOA/ePRO and External Data

External vendor data - including eCOA and ePRO - enhances precision in clinical trials, supports endpoint validation, ensures regulatory compliance and enables standardized reporting for safe, effective outcomes.



eCOA & ePRO

eCOA (Electronic Clinical Outcome Assessment) captures patient outcomes electronically. **ePRO (Electronic Patient-Reported Outcomes)** is a key subset where patients directly report health status without clinician interpretation.

Diaries

Wearables

Questionnaires

Strategic Value



Patient-Centric

Direct voice & experience capture



Real-Time Data

Immediate symptom tracking



Richer Endpoints

Better quality of life metrics



Decentralized

Supports remote trial designs



This Data plays a vital roles in collecting information of actual patient perspective in the clinical trials , however due to confidentiality and privacy concern original data is not easily available for further research by clinicians, statistician and academicians.



www.geninvo.com
North America | UK | Europe | India

Traditional Synthetic Data Approaches: Critical Limitations

Traditional synthetic data generation in clinical trials faces a "trilemma" of competing constraints: maintaining **privacy**, ensuring **data utility** and handling **operational complexity**.



Privacy & Re-identification

- ✗ Traditional methods rely on training models with **real patient data** (PHI/PII).
- ✗ High risk of "membership inference attacks" where specific patients can be identified.
- ✗ Regulatory hurdles (GDPR/HIPAA) for cross-border data sharing.



Missingness & Quality

- ⚠ Real-world data is "messy": missing values, irregular visits, and dropout.
- ⚠ Simple simulation often produces "too perfect" data that fails to test system robustness.
- ⚠ Inability to replicate complex correlations between domains



Non-Standardization

- ✗ DTA/DTS formats vary significantly across vendors.
- ✗ Manual mapping and reconciliation consume excessive time and resources.
- ✗ Lack of reusable test datasets leads to repetitive work for each study.



The Need: Realistic, regulation-aligned test data generated without accessing patient-level source data.



www.geninvo.com
North America | UK | Europe | India

Overcoming Barriers: A Python-Based Approach

Generating realistic external vendor datasets without ever touching patient-level information.

- **Leveraging Open-Source Python Tools**

Utilize DTA/DTS specification files as the sole data source —no patient information required

- **Creating Synthetic ePRO & eCOA Forms**

Generate compliant questionnaires and clinical outcome assessments from metadata alone

- **Incorporating Real-World Complexity**

Include missing values, edge cases, and "dirty data" patterns for authentic testing scenarios



Privacy-by-Design



Metadata-Based



Reproducibility



No PII/PHI Required



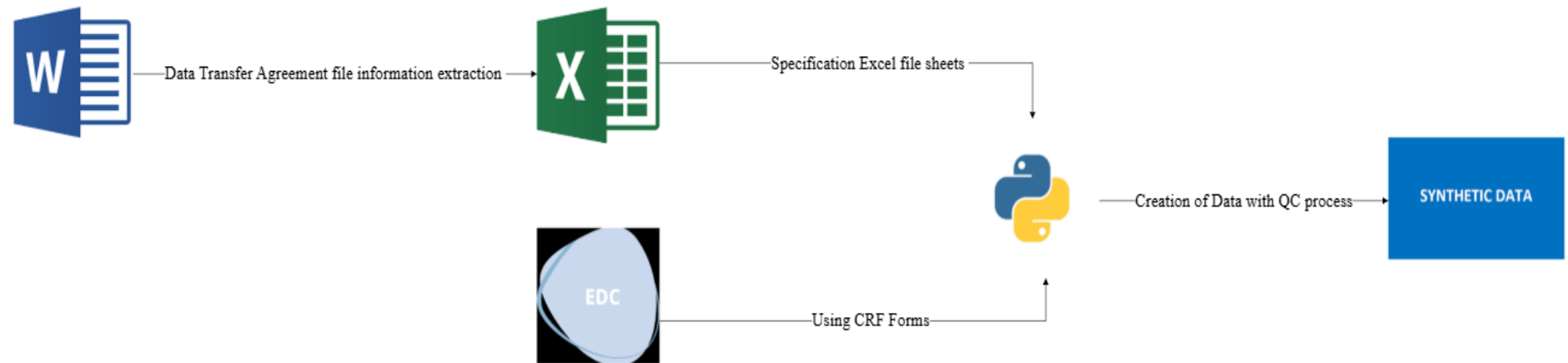
Regulation-Compliant



www.geninvo.com
North America | UK | Europe | India

Workflow: Steps for Synthetic Data Creation

1. Data Transfer Agreement file information extraction using ML models
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 ,LB(if present) file for visit, demographic and leading question information.
4. Algorithm processing and synthetic generation



Open-Source Python Stack

DATA CORE

NumPy

Numerical backbone enabling controlled random sampling and distribution modeling.

Sampling | Distributions

Pandas

Builds SDTM/ADaM-like tabular structures and applies metadata-driven rules.

DataFrames | Clinical Tables

NLP & SEMANTICS

Faker

Generates PII-safe subject identifiers, dates, sites, and demographics.

PII-Free | Identifiers

spaCy

Provides linguistic structure for synthetic narratives and medical text.

NER | Tokenization

QA & VALIDATION

Sentence Transformers

Ensures semantic similarity between real and synthetic clinical narratives.

Embeddings | Semantic QA

RapidFuzz

Validates realism using fuzzy matching and controlled variability checks.

Fuzzy Match | Quality



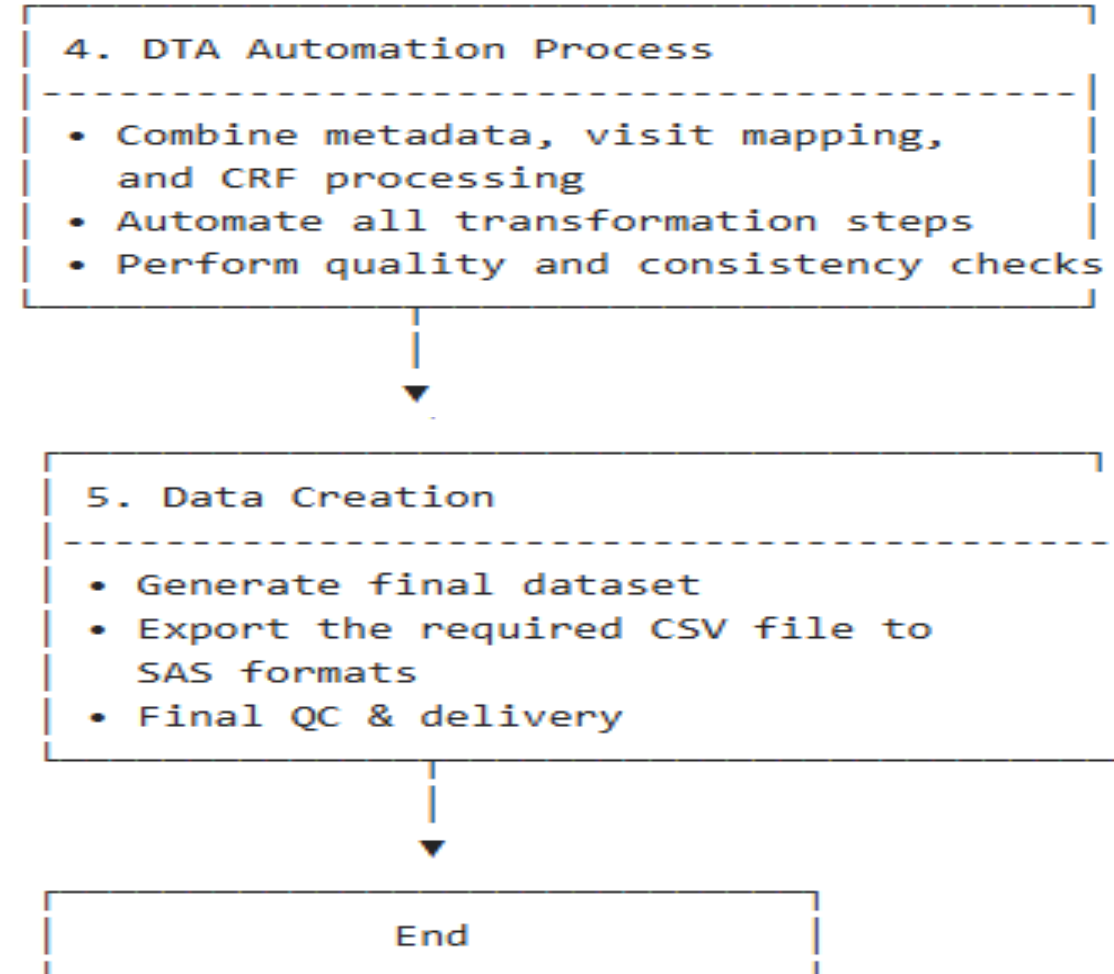
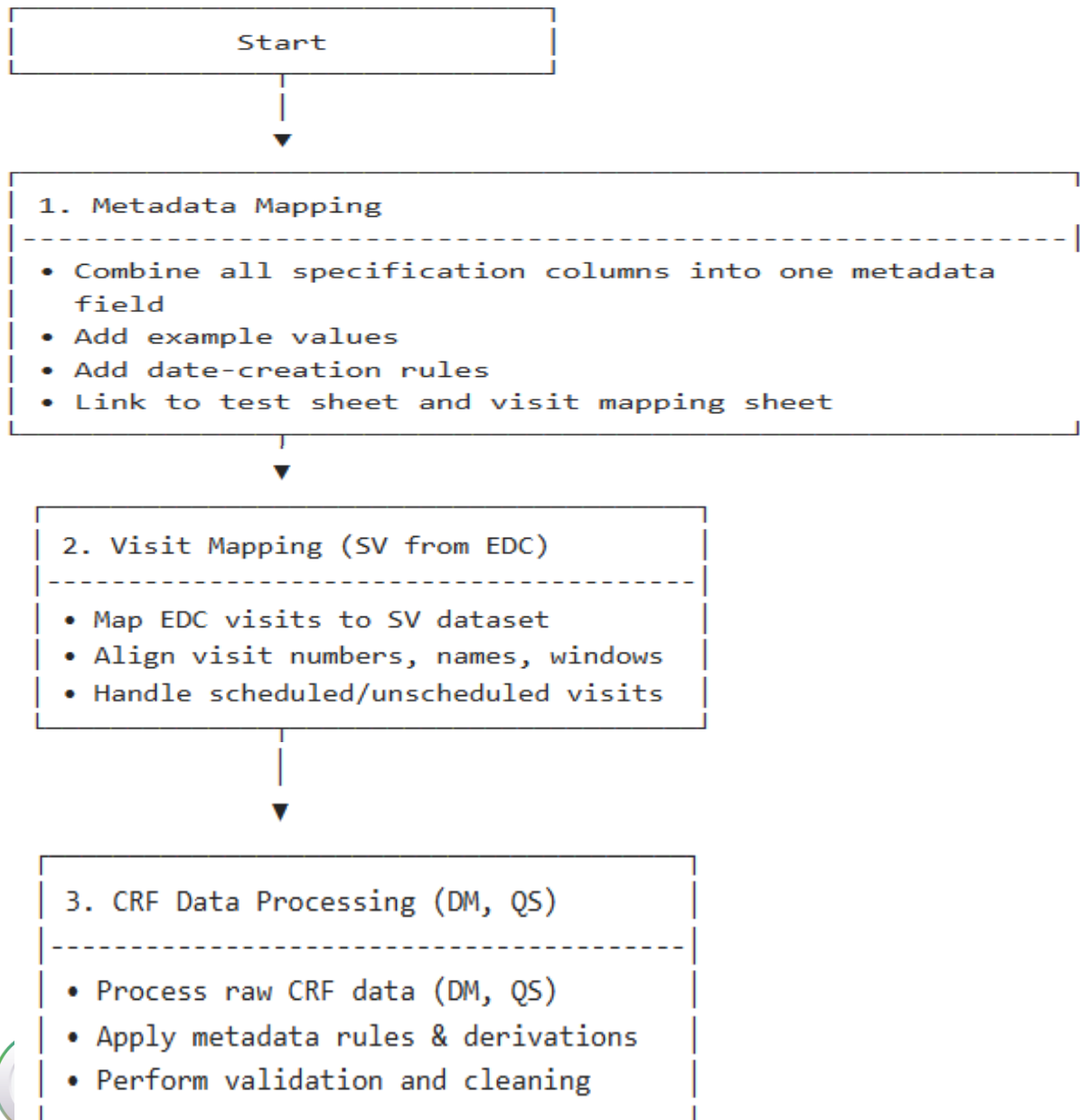
www.geninvo.com
North America | UK | Europe | India

Python Algorithm Processing:

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

```
1  Start
2  ↓
3  Metadata Mapping
4  ↓
5  Visit Mapping (SV from EDC)
6  ↓
7  CRF Data Processing
8  (DM, QS)
9  ↓
10 DTA Automation
11 ↓
12 Data Creation
13 ↓
14 End
```





Results

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

< >

QS_Data

+

Applications

Function	Example Use	Outcome
Data Management	Edit-checking Testing	Early error discovery
Programming	SDTM/ADaM mockup	Faster builds
Biostats	TFL dry runs	Reduced rework
AI/ML	Model training	Bias testing & scalability



Q&A



www.geninvo.com
North America | UK | Europe | India



Ideas
Delivered
Intuitively

Thank You!

+1-833-GEN-INVO (436-4686) | info@geninvo.com | www.geninvo.com

