Paper DH13

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

Jaskaran Singh Saini, GenInvo, Chandigarh, India
Hitesh Raval, GenInvo, Mumbai, India

## ABSTRACT

External vendor (third-party) data such as ePRO (Electronic Patient Reported Outcomes) and laboratory results etc. plays a critical role in clinical trials but introduces challenges, including data privacy concerns, missing values and non-standardization in Data Transfer Agreements (DTA) and Data Transfer Specifications (DTS). Traditional synthetic data generation often relies on real patient data to train machine learning models, posing risks of patient re-identification. This paper presents a novel approach to generating synthetic external vendor data for domains like QS (Questionnaires) and LB (Laboratory Test Results) without using patient-level data. By leveraging DTA/DTS metadata and open-source Python tools, this framework produces realistic, regulation-compliant datasets. It also incorporates synthetic Electronic Data Capture (EDC) forms, such as SV (Schedule of Visits) and DM (Demographics), to mirror real-world data complexity. This method enables secure testing, system validation, and fosters innovation in data simulation and integration, while ensuring compliance with evolving regulatory standards.

## INTRODUCTION

External vendor data plays an increasingly critical role in the successful execution of modern clinical trials. Data sources such as electronic Patient-Reported Outcomes (ePRO), electronic Clinical Outcome Assessments (eCOA), laboratory results and other third-party measurements enhance the precision of clinical insights, support endpoint validation and facilitate regulatory compliance, ultimately contributing to safe and effective trial outcomes. The integration of such data also enables standardized reporting across studies, supporting consistency and traceability throughout the clinical data lifecycle (CDISC, n.d.; PHUSE, 2023).

Among external data sources, eCOA and ePRO have emerged as key components of patient-centric trial designs. These technologies allow real-time digital capture of patient experiences - such as symptoms, functional status and quality of life - directly from participants without investigator interpretation. By reducing recall bias and transcription errors, eCOA and ePRO can improve data accuracy, increase operational efficiency and align with regulatory expectations for patient-focused drug development (Clario, 2024; Pharmaceutical Technology, 2024). However, the successful use of these tools requires robust technological infrastructure and strict adherence to validation and quality standards, including proper instrument handling and alignment with data standards for downstream tabulation and analysis (CDISC, n.d.; PHUSE, 2023).

Despite their advantages, obtaining reliable and analyzable eCOA and ePRO data remains challenging. Variability in devices and platforms, inconsistent patient engagement, missing or incomplete entries and complex data integration workflows across multiple vendors introduce significant operational and analytical complexities. These challenges are further compounded by non-standardized Data Transfer Agreements (DTAs) and Data Transfer Specifications (DTSs), which often differ by vendor and study, increasing the effort required for data integration, validation and downstream standardization (Minnakanti, 2023; Patil & Kumar, 2024).

Furthermore, while eCOA and ePRO data are essential for capturing the true patient perspective in clinical trials, the use of original patient data for secondary research, development and training purposes is highly restricted due to confidentiality, privacy and regulatory constraints. As a result, clinicians, statisticians and academicians often have limited access to realistic datasets for methodological research, system testing and skills development. This limitation highlights the need for alternative approaches that maintain data realism while protecting patient privacy, thereby enabling innovation without compromising ethical or regulatory obligations (NIH, n.d.; Arora et al., 2025).

## CHALLENGES

The increasing reliance on external vendor data, particularly eCOA and ePRO, introduces several challenges that can significantly affect data quality, timelines and regulatory readiness in clinical trials. One of the foremost challenges is adhering to regulatory compliance requirements. Regulatory authorities such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) emphasize the need for data integrity, traceability and validation of electronic systems used for outcome assessments. Ensuring compliance across multiple third-party vendors, each with distinct systems and processes, requires substantial oversight and robust governance frameworks (CDISC, n.d.; PHUSE, 2023).

Privacy and confidentiality concerns represent another major challenge. eCOA and ePRO data are inherently patient-centric and often contain sensitive health information. Strict regulations, including the Health Insurance Portability and Accountability Act (HIPAA) and global data protection laws, limit the use, sharing and reuse of original patient data. These restrictions, while essential for protecting patient rights, significantly constrain the availability of realistic datasets for research, system testing and methodological development by clinicians, statisticians and academicians (NIH, n.d.; Arora et al., 2025).

A persistent operational issue is the non-standardization of Data Transfer Agreements (DTAs) and Data Transfer Specifications (DTSs) across vendors. Agreement files frequently vary in structure, terminology, file formats and metadata definitions, even for similar data types such as ePRO questionnaires or laboratory results. This lack of standardization increases the complexity of data integration, mapping to CDISC SDTM domains and downstream analysis, often requiring study-specific custom programming and extensive reconciliation efforts (Minnakanti, 2023).

Data delays are also common when relying on external vendors. Delays may arise from scheduled transfer cycles, late data corrections, system outages or prolonged query resolution processes. Such delays can have downstream impacts on interim analyses, database locks and regulatory submission timelines, particularly in studies with time-critical endpoints or adaptive designs (PHUSE, 2023).

In addition, missing data and complex data integration workflows pose significant analytical challenges. In eCOA and ePRO data, missing entries may result from patient non-compliance, device malfunctions or usability issues. Integrating this data with EDC-collected information, such as demographics (DM) and visit schedules (SV), further increases complexity, especially when vendor data structures do not align cleanly with CDISC standards (Patil & Kumar, 2024).

These challenges often lead to costly rework, including repeated data transformations, remapping efforts and revalidation of analysis datasets. Rework not only increases programming and operational costs but also introduces additional risks to data consistency and traceability, particularly when changes occur late in the study lifecycle (Minnakanti, 2023; PHUSE, 2023).

Finally, a critical overarching challenge is balancing data utility with confidentiality. While realistic data are essential for testing, training and innovation, the use of original patient data is heavily restricted. Organizations must therefore find solutions that preserve the analytical utility and structural realism of eCOA and ePRO data while ensuring that patient privacy and regulatory obligations are fully maintained. This balance remains a central challenge in enabling efficient and compliant clinical data research and development (NIH, n.d.; Arora et al., 2025).

## LIMITATION OF TRADITIONAL SYNTHETIC DATA APPROCHES

Traditional synthetic data generation approaches have been increasingly explored in clinical research to address data availability and privacy constraints. However, when applied to complex external vendor data such as eCOA and ePRO, these approaches present several limitations that restrict their practical utility and regulatory acceptability.

One critical limitation is the risk to data privacy. Many traditional synthetic data techniques rely on training statistical or machine-learning models directly on original patient-level data to learn underlying distributions and relationships. Without robust privacy-preserving mechanisms and formal risk assessments, these models may inadvertently retain or reproduce sensitive patterns, increasing the risk of patient reidentification. This concern is particularly significant in high-dimensional clinical datasets where combinations of attributes can enable identity inference, even in ostensibly de-identified datasets (NIH, n.d.; Arora et al., 2025).

Regulatory and validation challenges further limit the use of conventional synthetic data methods. Regulatory agencies require transparency, traceability and validation of data generation processes, particularly when synthetic data are used to support decision-making, system testing or methodological research. Many traditional approaches operate as “black box” models, offering limited explainability, weak documentation of assumptions and insufficient controls for reproducibility, all of which complicate qualification and acceptance in regulated clinical environments (CDISC, n.d.; PHUSE, 2023).

Another key challenge is the limited ability of traditional models to capture complex clinical relationships. Clinical trial data, particularly eCOA and ePRO, are often longitudinal, hierarchical and highly contextual, incorporating subject-level, visit-level and item-level dependencies. Conventional synthetic data methods may struggle to preserve these relationships accurately, leading to unrealistic temporal patterns, loss of visit structure or invalid correlations between domains such as demographics, visit schedules and outcome assessments (Patil & Kumar, 2024).

Traditional approaches are also associated with high computational costs, especially when applied to large, high-dimensional datasets. Advanced generative models, including deep learning-based methods, often require extensive computational resources, long training times and specialized hardware. These requirements increase operational costs and restrict adoption in environments where lightweight, repeatable and transparent data generation processes are preferred (Del Gobbo, 2025).

A well-recognized limitation is the poor modeling of rare adverse events and rare disease signals. Rare clinical outcomes are under-represented in training data. Traditional synthetic data models tend to smooth distributions and optimize for common patterns, resulting in the dilution or complete loss of rare but clinically meaningful signals. This limitation reduces the utility of such data for safety testing, edge-case validation and methodological research in rare disease contexts (Del Gobbo, 2025; Arora et al., 2025).

In addition, traditional synthetic data methods can lead to bias preservation and amplification. If underlying biases exist in the source data - such as demographic imbalances, differential completion rates or systematic missingness - synthetic data models may replicate or even amplify these biases. Without explicit bias detection and correction mechanisms, the resulting synthetic datasets may reinforce existing inequities and misrepresent patient populations (Nature Digital Medicine, 2023).

Scalability challenges in high-dimensional settings further constrain traditional approaches. As the number of variables, domains and longitudinal dimensions increases, many models experience degradation in data quality, stability and consistency. This is particularly problematic for clinical datasets that integrate multi-domain data such as DM, SV, QS, and LB, each with domain-specific constraints and controlled terminology requirements (CDISC, n.d.).

Finally, traditional synthetic data solutions often demonstrate limited transportability across studies, therapeutic areas or vendor ecosystems. Models trained on a specific dataset may not generalize well to new studies, instruments or vendor formats without complete retraining. This lack of portability reduces reusability and undermines the goal of creating scalable, vendor-agnostic synthetic data solutions for clinical trial research and development (PHUSE, 2023).

## OVERCOMING BARRIERS: A PYTHON BASED APPROACH

To address the limitations associated with traditional synthetic data approaches and the operational challenges of using real patient data, a structured metadata-driven strategy for generating external vendor datasets is required. This paper proposes an approach that focuses on producing realistic, standards-aligned synthetic data for ePRO, eCOA and related electronic data capture (EDC) domains, while protecting patient confidentiality.

A key aspect of overcoming these barriers is leveraging open-source Python tools in conjunction with Data Transfer Agreement (DTA) and Data Transfer Specification (DTS) files. DTS documents define variable-level metadata, data structures, controlled terminology, file formats and transfer frequencies expected from external vendors. By programmatically interpreting these specification files, synthetic data can be generated to closely reflect vendor specific structures without deriving information from real patient records. Open-source Python libraries provide flexible, transparent and reproducible mechanisms for implementing such rule-based and constraint-driven data generation workflows, making them well suited for use in regulated clinical environments (Minnakanti, 2023; Del Gobbo, 2025).

This approach enables the creation of synthetic ePRO, eCOA and EDC forms without the use of real patient data. Domains such as Questionnaires (QS) for ePRO and eCOA and supporting EDC domains including Demographics (DM) and Subject Visits (SV), can be constructed using CDISC standards, controlled terminology and protocol-defined visit schedules. By separating data structure and business logic from patient-specific content, the resulting datasets preserve analytical realism while eliminating privacy risks associated with handling protected health information (CDISC, n.d.; NIH, n.d.). Such datasets can be safely shared across programming, validation and training teams to support development and quality assurance activities.

An additional strength of this framework is the intentional incorporation of real-world data characteristics, including missing values, edge cases and so-called “dirty data.” In operational clinical trials, ePRO and eCOA data frequently contain incomplete entries due to patient non-compliance, device issues or visit deviations. Similarly, external vendor datasets may include delayed records, inconsistent units or unexpected value combinations. By explicitly modeling these behaviors within synthetic datasets, the generated data more accurately reflects the conditions encountered in real studies. This allows teams to proactively test data integration pipelines, validation checks and analysis logic under realistic conditions, improving robustness and reducing costly downstream rework (Patil & Kumar, 2024; PHUSE, 2023).

Collectively, this metadata-driven, Python-based approach provides a practical pathway for overcoming privacy, regulatory and operational barriers to external vendor data usage. It enables the generation of high-fidelity synthetic datasets that are fit for purpose, standards-compliant and safe for widespread use, thereby supporting innovation and efficiency across the clinical trial data lifecycle.

## OPEN-SOURCE PYTHON STACK

The synthetic data generation framework leverages a modular open-source Python stack to simulate structurally compliant and statistically realistic clinical trial datasets. Each component contributes to a distinct layer of the synthetic data creation lifecycle, including numerical simulation, structural assembly, semantic enrichment and quality validation.

### NUMPY

NumPy serves as the numerical foundation of the framework. It enables controlled random sampling from parametric and non-parametric statistical distributions to simulate continuous and categorical clinical variables such as laboratory values, vital signs, adverse event counts etc. Seeded randomization ensures reproducibility, which is essential for pipeline validation and repeatable system testing. Additionally, NumPy supports correlation modeling across variables, allowing synthetic datasets to preserve realistic inter-variable relationships observed in clinical studies.

### PANDAS

Pandas provides the structural backbone for assembling SDTM and ADaM-like tabular datasets. It facilitates metadata-driven transformations, derivations and rule-based logic consistent with CDISC standards. Using DataFrame operations, domain-level datasets (e.g., DM, AE, LB) are constructed with appropriate keys, relationships and controlled terminology alignment. Pandas also enable structural validation and dataset integrity checks prior to downstream integration or analytical testing.

### FAKER

For generating realistic contextual identifiers, Faker is used to create fictitious subject IDs, site identifiers, dates, and demographic attributes. This approach ensures that synthetic datasets contain plausible operational metadata while eliminating any reference to real individuals. By design, no actual subject-level records or identifiers are reused, thereby supporting privacy protection and minimizing re-identification risk in non-production environments.

### SPACY

To support synthetic narrative generation, spaCy provides natural language processing capabilities, including tokenization, linguistic structuring and named entity recognition. These features enable the creation of coherent medical narratives such as medical histories or adverse event descriptions that follow domain-appropriate terminology patterns. This enhances the realism of text-based fields while maintaining separation from original patient data.

### SENTENCE TRANSFORMER & RAPID FUZZ

Sentence Transformers generate embedding-based representations of textual data, enabling quantitative assessment of semantic similarity between real and synthetic narratives without replicating source content. RapidFuzz supports fuzzy matching and controlled terminology validation to ensure alignment with CDISC standards and sponsor-defined dictionaries. Together, these tools provide an additional validation layer to confirm structural consistency, semantic coherence and controlled variability in the synthetic outputs.

Collectively, this open-source stack enables the generation of synthetic clinical datasets that replicate structural, statistical and semantic characteristics of real-world trial data while maintaining privacy-by-design principles. The modular architecture also allows extensibility and adaptability to evolving standards and study-specific requirements.

## WORKFLOW: STEPS FOR SYNTHETIC DATA CREATION

A structured and transparent workflow is essential for generating high-quality synthetic external vendor data that is realistic, standard aligned and suitable for use in regulated clinical research settings. The proposed workflow integrates transfer agreement-level metadata, electronic data capture (EDC) information and algorithmic processing to create synthetic ePRO, eCOA related datasets without relying on real patient data. (Figure-1)

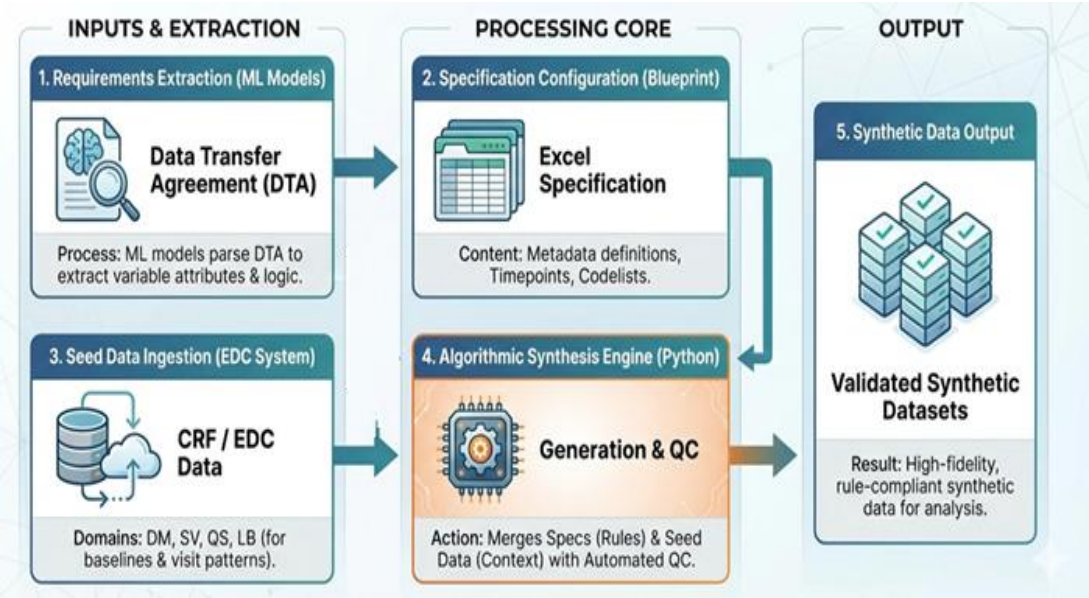



Figure-1

1. **Extraction of Data Transfer Agreement Information**

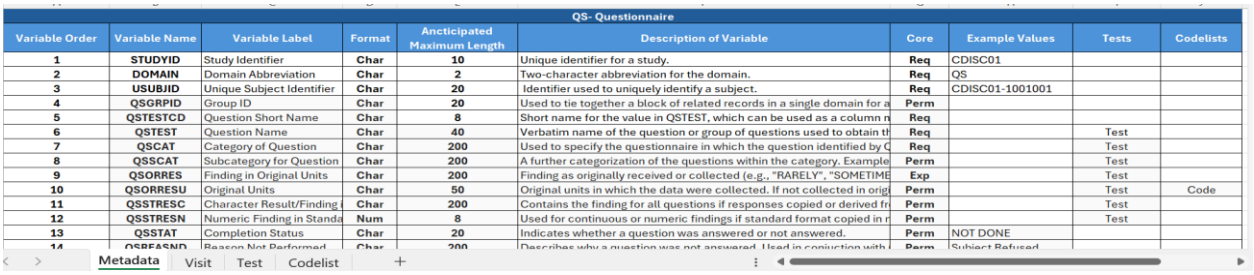
   The first step in the workflow involves extracting relevant information from Data Transfer Agreement (DTA) and Data Transfer Specification (DTS) files. These documents typically contain detailed definitions of dataset structures, variable names, data types, controlled terminology, transfer frequency and validation rules specified by external vendors. Given the variability in format and structure across vendors, machine learning (ML)-assisted techniques are being applied to automate the identification and classification of key metadata elements within these documents. This includes parsing semi-structured text, tables and PDF-based specifications to systematically capture data requirements in a machine-readable form.

   Automating DTA/DTS information extraction helps reduce manual errors, improves consistency in metadata interpretation and enables scalable reuse of specifications across studies and vendors. Importantly, this step focuses solely on agreement-level and specification-level content and does not involve patient-level data, thereby maintaining compliance with privacy and confidentiality regulations.

2. **Creation of a Structured Specification File**

   Following metadata extraction, a centralized specification file, typically in Excel format, is created to serve as the blueprint for synthetic data generation. This specification file contains multiple structured sheets, such as Metadata, Timepoints, Codelist, Tests and Domain Mapping. Each sheet captures a distinct aspect of the expected data structure, including visit schedules, assessment windows, allowable values and relationships between domains.

   This specification-driven artifact acts as a single source of truth, aligning external vendor expectations with CDISC standards such as SDTM and the Questionnaires, Ratings and Scales (QRS) supplements. By formalizing requirements in a structured and auditable manner, the specification file supports traceability, reproducibility and validation - key considerations in regulated clinical environments. End user QC is performed to validate the specification file aligns with the DTA/DTS files before moving to next step. (Figure-2,3)

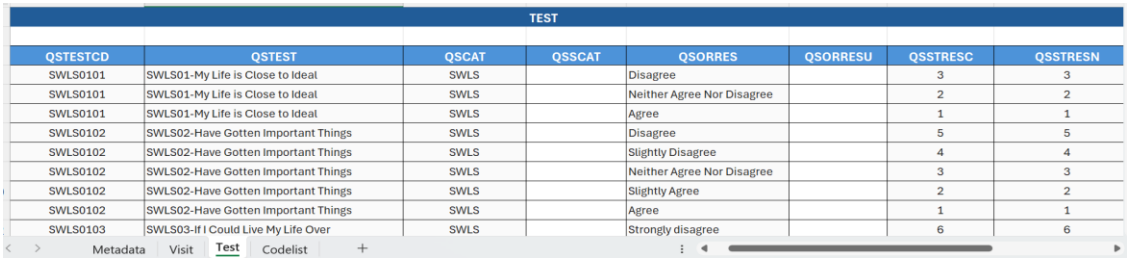


QS- Questionnaire

| 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 n | 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 C | 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 origi | Perm | | Test | Code |
| 11 | QSSTRESC | Character Result/Finding | Char | 200 | Contains the finding for all questions if responses copied or derived fr | Perm | | Test | |
| 12 | QSSTRESN | Numeric Finding in Standa | Num | 8 | Used for continuous or numeric findings if standard format copied in n | Perm | | Test | |
| 13 | QSSTAT | Completion Status | Char | 20 | Indicates whether a question was answered or not answered. | Perm | NOT DONE | | |
| 14 | QSREASND | Reason Not Performed | Char | 200 | Describes why a question was not answered. Used in conjuction with | Perm | Subject Refused | | |

Metadata | Visit | Test | Codelist

Figure-2



TEST

| QSTESTCD | QSTEST | QSCAT | QSSCAT | QSORRES | QSORRESU | QSSTRESC | 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 |

Metadata | Visit | Test | Codelist

Figure-3

## 3. Integration of EDC (CRF) Domain Information

In the next step, relevant EDC/CRF derived domain information is incorporated to provide clinical context and structural coherence. Core domains such as Subject Visits (SV), Demographics (DM) and Questionnaires (QS) when applicable - are used to define visit structures, subject characteristics, and leading question information. These domains establish the backbone for aligning synthetic external vendor data with study design elements such as visit schedules, eligibility criteria and assessment timing.

Integrating EDC derived structures ensures that synthetic ePRO and eCOA data are not generated in isolation but are logically consistent with the overall study framework. This step also facilitates realistic modeling of longitudinal relationships, such as subject level repeated measures across visits, while remaining independent of real patient records.

## 4. Algorithmic Processing and Synthetic Data Generation

The final step involves algorithmic processing and synthetic dataset generation using open-source Python-based methods. Rule-based logic and constraint-driven algorithms are applied to the specification file and EDC domain structures to generate synthetic datasets that conform to expected formats and controlled terminology. Where appropriate statistical algorithms, machine learning–based components are used to introduce realistic variability, correlations and distributions, while ensuring that no identifiable patient information is introduced.

This step also allows for the intentional incorporation of real-world operational characteristics, such as missing values, delayed records, edge cases and inconsistent entries, thereby producing datasets that reflect real trial conditions. The steps for the Algorithmic Processing at each step with details are mentioned in Figure 4.

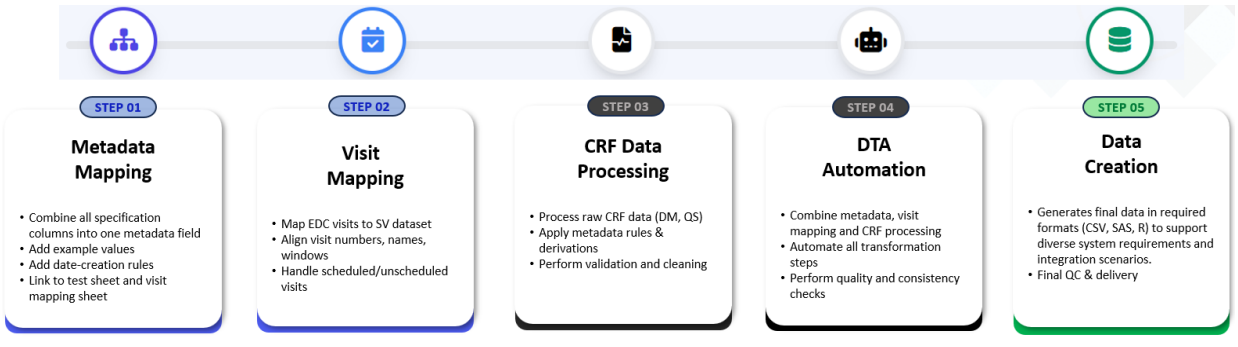



Figure-4 Algorithmic Processing Workflow

## RESULT

The resulting synthetic outputs can be used for pipeline testing, integration validation, programmer training and methodological research, all while maintaining regulatory compliance and protecting patient confidentiality. Regulatory compliance is supported by generating datasets that replicate structural and statistical characteristics of clinical trial data without retaining or reconstructing identifiable subject-level records. These datasets are intended for non-submission purposes and preserve patient confidentiality by design.

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | STUDYID | DOMAIN | USUBJID | QSGRPID | QSTESTCD | QSTEST | 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

Figure-5 – Final dataset in .csv format

## APPLICATION ACROSS FUNCTIONS

The generated synthetic external vendor datasets for ePRO, eCOA and supporting EDC domains have broad applicability across multiple functional areas within the clinical trial lifecycle. By closely mimicking real-world vendor data structures while remaining free of patient-identifiable information, these datasets enable teams to perform critical development, testing and innovation activities more efficiently and safely.

### Data Management

From a Data Management perspective, synthetic datasets are highly effective for edit-check and validation testing. Realistic ePRO and eCOA data patterns - including missing responses, out-of-range values and inconsistent entries- allow data managers to proactively test edit checks, data validation rules and reconciliation logic before live data transfers begin. This enables early error discovery, reducing the likelihood of downstream discrepancies during production data loads and minimizing the volume of late-stage data queries.

### Statistical Programming

For Statistical Programming, synthetic vendor data provides a reliable foundation for SDTM and ADaM mockups. Programmers can use these datasets to develop and validate mapping logic for domains such as QS, LB, DM, and SV, as well as to rehearse derivations for analysis-ready datasets (e.g., ADQS). Because the synthetic data are specification-driven and CDISC-aligned, programs developed against them are more robust and reusable, resulting in faster dataset builds once live data become available.

### Biostatistics

Within Biostatistics, synthetic data support Tables, Listings, and Figures (TLF) dry runs early in the study lifecycle. Biostatisticians can test analysis specifications, scoring algorithms (e.g., total and subscale scores for questionnaires), and reporting layouts without waiting for real patient data. This early validation of analysis logic and outputs helps identify design or specification gaps sooner, leading to reduced rework and fewer late-stage changes during database lock and reporting phases.

### Artificial Intelligence and Machine Learning

Synthetic external vendor data also enables important use cases in Artificial Intelligence and Machine Learning (AI/ML). These datasets can be safely used for model training, bias evaluation and scalability testing without exposing sensitive patient information. By incorporating realistic distributions, missingness patterns and edge cases, the synthetic data allow teams to assess model behavior under real-world conditions, test for potential biases and evaluate performance across varying dataset sizes. This supports responsible AI development while maintaining strict privacy and regulatory boundaries.

## KEY UTILITIES & BENEFITS

The synthetic data created has lot of potential utility, some of the key utilities and benefits are mentioned below.

### Data Privacy and Security

Because the datasets are generated from DTA/DTS specifications and standards metadata - not from identifiable records - they can be shared broadly across teams without invoking PHI handling or authorization workflows. This materially reduces legal and operational risk while aligning with de-identification principles under HIPAA and best-practice guidance for synthetic health data.

### Scenario Simulation

Teams can safely simulate operational realities - missed ePRO entries, device drop-outs, late vendor transfers, unit mismatches - before first patient in. “What-if” scenarios stress pipelines, reconciliation rules and visit window logic across QS/LB/DM/SV, enabling proactive control-plan design and smoother go-lives.

### Quality Control and Testing

Specification-driven datasets support early QC of edit checks, SDTM mappings and ADaM derivations. Intentional inclusion of edge cases and “dirty data” enables robust negative testing; repeated runs verify reproducibility and traceability consistent with CDISC expectations.

### Algorithm Training and Validation

For analytics and AI/ML, these synthetic datasets enable model prototyping, bias evaluation and scalability testing without exposing patient identifiers. By encoding realistic distributions, correlations and missingness patterns, teams can evaluate performance and fairness while keeping the development environment privacy-preserving.

### Improving Data Privacy

The approach is privacy-by-construction. No original patient rows are ingested. Where statistical generators are used, they are constrained by specifications and reviewed with privacy metrics (e.g., uniqueness/neighbor distance) to minimize re-identification risk - supporting defensible governance narratives.

### Enhancing Data Utility

Utility is maximized by grounding synthesis in CDISC domain shapes, QRS supplements and Controlled Terminology, ensuring the outputs are directly usable for mapping rehearsals, TLF dry runs and submission-readiness checks. This balance fit-for-purpose utility with formal privacy safeguards - is central to accelerating development while maintaining compliance.

## CONCLUSION

This paper demonstrated a specification-driven, Python-based approach to generate realistic, standards-aligned synthetic external vendor data - notably for QS and LB, with contextual DM and SV - without using patient-level records. By anchoring synthesis in DTA/DTS metadata, CDISC SDTM/QRS structures and intentional modeling of operational realities (missingness, delays, edge cases), the method delivers datasets that are fit for testing, validation and analysis rehearsals while protecting privacy. The workflow - DTA/DTS extraction, structured specification build, EDC integration and algorithmic generation - enables earlier error discovery, faster SDTM/ADaM builds, reduced rework and safer AI/ML prototyping. Although instrument licensing, vendor heterogeneity and ongoing privacy/bias evaluation remain important considerations, the framework offers an immediately actionable path to balance utility with confidentiality. Looking ahead, standardizing quality/privacy metrics and automating metadata deliverables (e.g., Define-XML/VLM) will further streamline adoption across studies and vendors.

## REFERENCES


1. **CDISC.** (n.d.). *Standards*. Clinical Data Interchange Standards Consortium. https://www.cdisc.org/standards
2. **PHUSE.** (2023). *PHUSE US Connect 2023—Paper & pre-recording guidelines*. https://cdn.eventsforce.net/files/ef-xpmm6vy566ze/website/18/paper_guidelines_us_connect_23.pdf
3. **Minnakanti, M.** (2023). *Reviewing external vendor data transfer specifications from programming standpoint (Paper DH12)*. PHUSE US Connect. https://www.lexjansen.com/phuse-us/2023/dh/PAP_DH12.pdf
4. **Patil, V., & Kumar, M.** (2024). *Streamlining patient-reported outcome (PRO) data standardization & analysis (Paper DS-398)*. PharmaSUG 2024. https://www.lexjansen.com/pharmasug/2024/DS/PharmaSUG-2024-DS-398.pdf
5. **U.S. NIH.** (n.d.). *HIPAA Privacy Rule and its impact on research*. https://privacyruleandresearch.nih.gov/pr_08.asp

6. **Arora, A., Wagner, S. K., Carpenter, R., Jena, R., & Keane, P. A.** (2025). The urgent need to accelerate synthetic data privacy frameworks for medical research. *The Lancet Digital Health, 7*(2), e157–e160. https://www.thelancet.com/journals/landig/article/PIIS2589-7500%2824%2900196-1/fulltext
7. **Clario.** (2024). *What is eCOA? Understanding its power and significance in clinical trials*. https://clario.com/resources/articles/what-is-ecoa-and-how-does-it-improve-clinical-trial-data-quality/
8. **Pharmaceutical Technology (Datacubed Health partner content).** (2024, May 1). *ePRO vs eCOA: Understanding the nuances in clinical trials*. https://www.pharmaceutical-technology.com/sponsored/epro-vs-ecoa-understanding-the-nuances-in-clinical-trials/
9. **CDISC.** (n.d.). *QRS—Questionnaires, Ratings, and Scales*. https://www.cdisc.org/standards/foundational/qrs
10. **CDISC / NCI EVS.** (2025, Sept 26). *CDISC Controlled Terminology (SDTM)*. https://evs.nci.nih.gov/ftp1/CDISC/SDTM/SDTM%20Terminology.html
11. **Marneni, M., & Olexovitch, S.** (2023). *Standardizing SDTM Laboratory (LB) Programming (Paper SI09)*. PHUSE US. https://www.lexjansen.com/phuse-us/2023/si/PAP_SI09.pdf
12. **Guillaudeux, M., Rousseau, O., Petot, J., et al.** (2023). Patient-centric synthetic data generation: No reason to risk re-identification in biomedical data analysis. *npj Digital Medicine, 6*(37). https://www.nature.com/articles/s41746-023-00771-5
13. **Del Gobbo, C.** (2025). *A comparative study of open-source libraries for synthetic tabular data generation: SDV vs SynthCity*. arXiv. https://arxiv.org/abs/2506.17847
14. **SDV (Synthetic Data Vault) Documentation.** (n.d.). *Welcome to the SDV!* https://docs.sdv.dev/sdv

## ACKNOWLEDGMENTS

We would like to express our heartfelt gratitude to Geninvo Technology Pvt Ltd and PHUSE for providing the opportunity to contribute to this paper, as well as for their continuous support and encouragement throughout the research process. We also extend our sincere appreciation to all those who contributed to the completion of this paper, with special thanks to our colleagues for their valuable feedback and support.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Jaskaran Singh Saini
Company: Geninvo Technologies Pvt. Ltd.
Address: 1408 East Empire Street 61701 Bloomington (Illinois), United States.
Work Phone: +1 309-807-5006
Email: jaskaran.s206@geninvo.com
Website: GENINVO is the go-to partner for life science industry.

Author Name: Hitesh Raval
Company: Geninvo Technologies Pvt. Ltd.
Address: 1408 East Empire Street 61701 Bloomington (Illinois), United States.
Work Phone: +1 309-807-5006
Email: hitesh.r263@geninvo.com
Website: GENINVO is the go-to partner for life science industry.