Paper AD08

USDM Synthetic Data Generation

Stuart Malcolm, Veramed, London, UK

ABSTRACT

"We cannot start until we get some data" is the reason for many a delay during the early stages of a biometrics project, and programming to the data during early stages of the project leads to rework later when the volume of data increases.

Current approaches to generating either random test data or anonymized test data have limitations and don't meet both key requirements - that test data is realistic and easy to generate.

This presentation outlines an alternative approach to generation of realistic test data that uses the digital protocol USDM standard to generate a synthetic population and test scenarios for each subject based on the protocol design and schedule of assessments defined in the protocol, resulting in test data that is both realistic and easy to generate.

INTRODUCTION

The move to develop standardized digital protocol formats is part of our industry digital transformation – creating new opportunities for automation, reuse and efficiency throughout the clinical trial lifecycle.
However, a persistent challenge remains: the need for realistic, protocol-driven test data early in biometrics project is needed to ensure these metadata-driven processes are fit-for-purpose when deployed on real, live data.
This paper introduces an approach for generating synthetic clinical trial data using the Unified Study Definitions Model (USDM), enabling earlier quality assurance, reducing rework, and supporting regulatory readiness.

WHY DO WE NEED SYNTHETIC DATA?

Delays in clinical trial programming and data science workflows are frequently caused by the lack of suitable test data at the outset of a project. Traditional approaches, such as using random or placeholder data, often fail to capture the complexity and protocol-specific logic required for robust system development and validation. This can result in significant rework when real data becomes available, as early solutions may not align with the actual study design or subject pathways. Additionally, relying on anonymized data introduces compliance risks and is often impractical due to privacy regulations and limited early access.

REVIEW OF EXISTING SOLUTIONS

Several tools currently exist to generate synthetic data – popular CDISC-related projects are summaries in the table below. All current solutions are based on either generating random data or scrambling existing data.

Project / Tool	Standards	Capability	License
cdiscdataset.com	EDC, CRF, SDTM, SEND, ADaM, TLF	number of subjects, TA, Treatment Arms only. No Population control.	Proprietary
pharmaverseraw	EDC-agnostic. Mix of CDASH/Non-CDASH.	based on the CDISC Pilot Study SDTM	Apache 2
pharmaversesdtm	SDTM	Mix of TA-agnostic (DM, VS, EG, RS...) and TA-specific (oe_ophtha, rs_onco, rs_onco_irecist...)	Apache 2
random.cdisc.data	ADaM	Generate wide range of 'realistic' ADaM domains programatically	Apache 2
raw.synthetic.data	n/a	Proposed COSA project. Not released.	n/a
Ssdtm	SDTM	Synthetic data generators in Python	MIT
Synthpop	Any	R Synthetic data created from original dataframe using statistical distributions	GPL 3

Current solutions are generally adequate for generating random test data, which is useful for basic unit testing and verifying individual components in isolation. However, these approaches fall short when it comes to quality assurance activities that require data consistency with the trial design.

For example, scenarios where data points must be logically aligned across multiple dataframes—such as ensuring subject demographics, visit schedules, and assessment results are coherent—cannot be reliably simulated with random data alone.

As a result, these tools are insufficient for comprehensive system validation and protocol-driven testing.

For these scenarios what we need is not synthetic data generation – but rather synthetic study generation.

ARCHITECTURAL DESIGN

The intended use of Synthetic Study Data is part of a move to prospective quality assurance rather than retrospective quality control – in this context, a key requirement is that we generate a description of the synthetic data in addition to just the synthetic data, so that we can determine if results of processing the data are correct.

To address these limitations, a modular architecture has been developed, centered around the concept of a "subject journey" specification. In this approach, an intermediate narrative is generated for each synthetic subject, describing their path through the study in terms of arms, visits, interventions, and assessments.

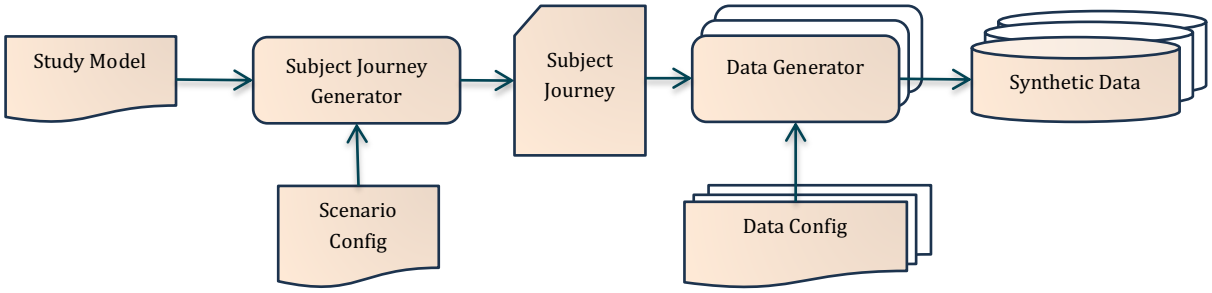

Figure 1 Modular architecture centered around the Subject Journey

This subject journey is independent of any specific data standard, allowing the same specification to be used for generating EDC extracts, raw datasets, or SDTM-compliant outputs.

By decoupling the journey specification from the final data format, this architecture supports scenario-based testing, easier updates, and greater reusability across different systems and use cases.

KEY ELEMENTS OF USDM

Unified Study Definitions Model (USDM) provides a standard for protocol-driven synthetic data generation.

USDM provides a comprehensive, machine-readable framework that captures all essential elements of a clinical trial's design, enabling the creation of realistic and standards-aligned synthetic datasets.

In this section we consider the key elements of USDM that we need for synthetic data generation.

STUDY DESIGN

The StudyDesign class is the container for a single design within a study definition. It can be either an observational study design or an interventional study design, and both classes include references to study timelines, populations and design elements like arms, epochs, and encounters.

The StudyDesign also provides slots for key parameters such as therapeutic area, study phase, and study type.

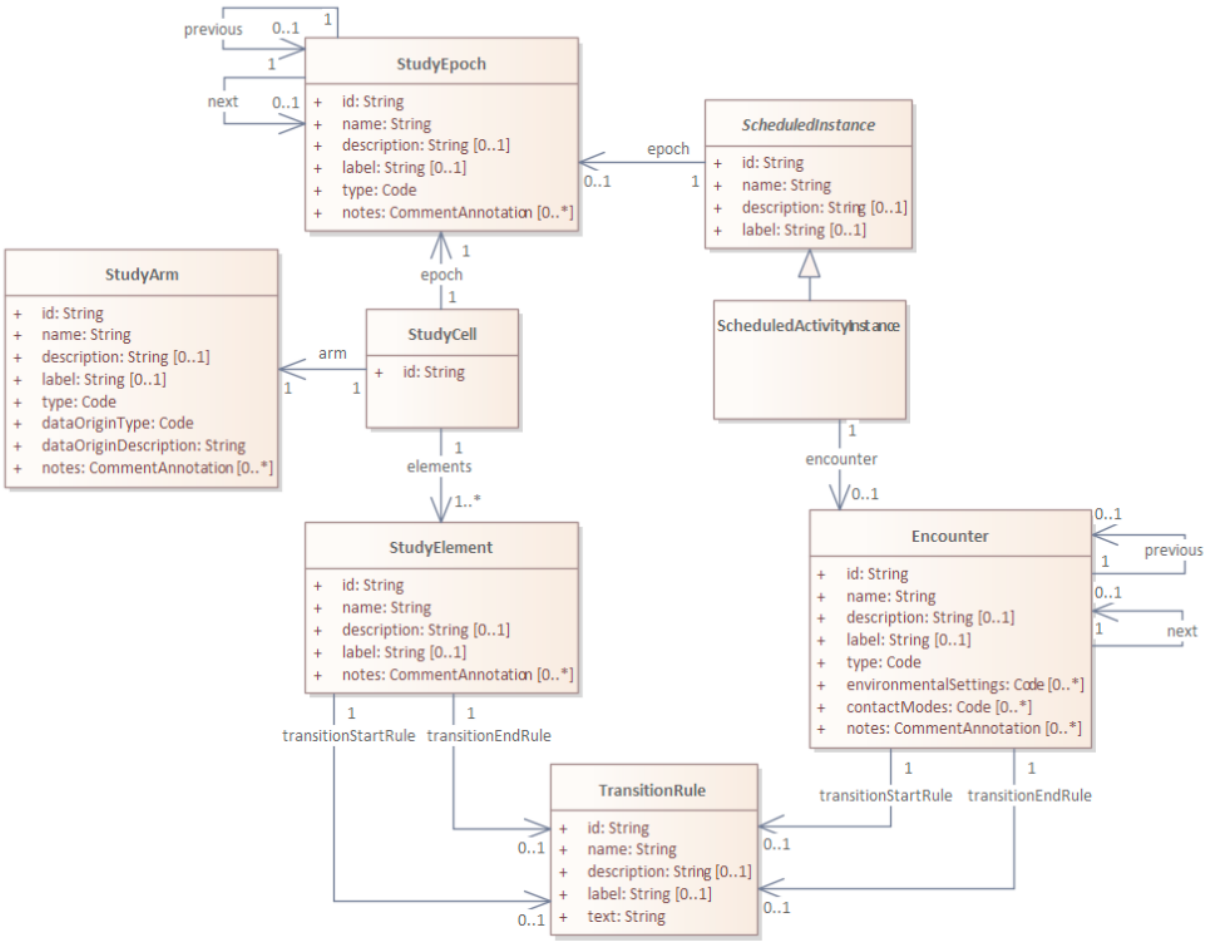

Figure 2 USDM Study Design based on Arms, Epochs is aligned with SDTM

The high-level study design based on arms and epochs is defined using the StudyArm, StudyEpoch, StudyCell, and StudyElement classes. The manner in which the classes are used follows the CDISC SDTM. Epochs are related to the study encounters (a more generic term for visits) via ScheduledInstances that form a ScheduleTimeline .

Modelling the study design this way, and linking the design to the population and SoA, allows us to “walk” a synthetic subject through the trial, which is the basis of the synthetic study data generation workflow (see later section)

POPULATIONS

Populations and cohorts represent groups or subgroups of subjects participating in a study. The PopulationDefinition class provides a general framework for defining these groups, including references to applicable eligibility criteria. StudyDesignPopulation captures population details for a specific study design, while StudyCohort describes subpopulations that may differ in treatment, assessment, or analysis based on subject characteristics. Attributes such as planned completion or enrolment numbers, planned age, and planned sex can be specified at either the population or cohort level, and additional cohort characteristics can be defined as needed.

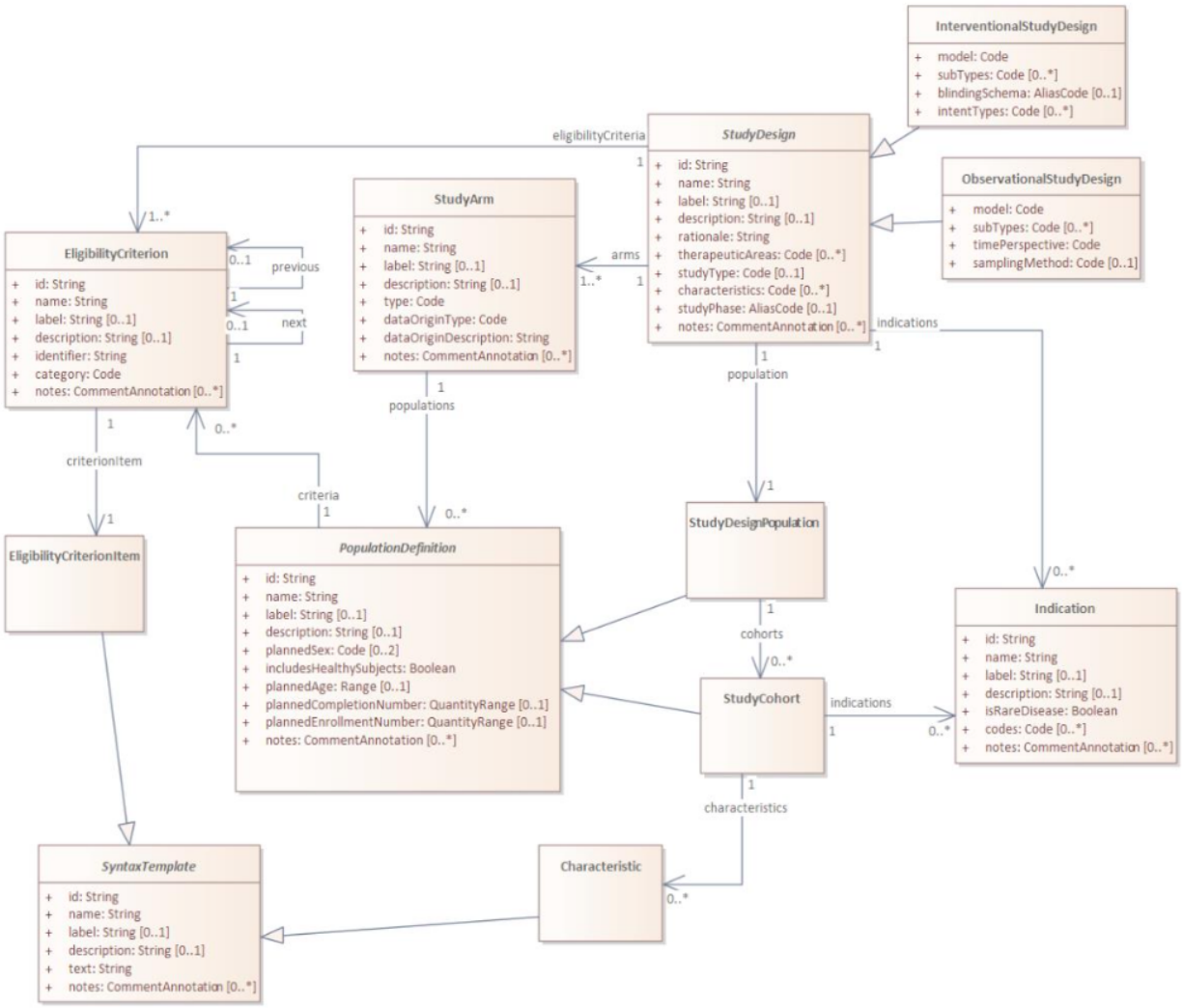

Figure 3 USDM Population, Cohort and Eligibility classes

Eligibility criteria are structured items that determine subject inclusion or exclusion and inherit from a common Syntax Template, allowing reference to any item within the USDM. The ordering of eligibility criteria can be specified for clear presentation, and each criterion can be mapped to standard domains such as SDTM Trial Inclusion/Exclusion Criteria. Study indications, including rare disease status, are also defined and can be linked to the overall population or specific cohorts. Notes and codes may be added to provide further context or alignment with internal or external standards.

SCHEDULE OF ACTIVITIES

One of the key aspects of a study design is the timing of encounters (visits) and the activities to be performed within those encounters. The USDM includes a mechanism for building timelines that can be reused within a study and, given external library management, across studies. The corresponding classes and attributes are shown in the following UML diagram.

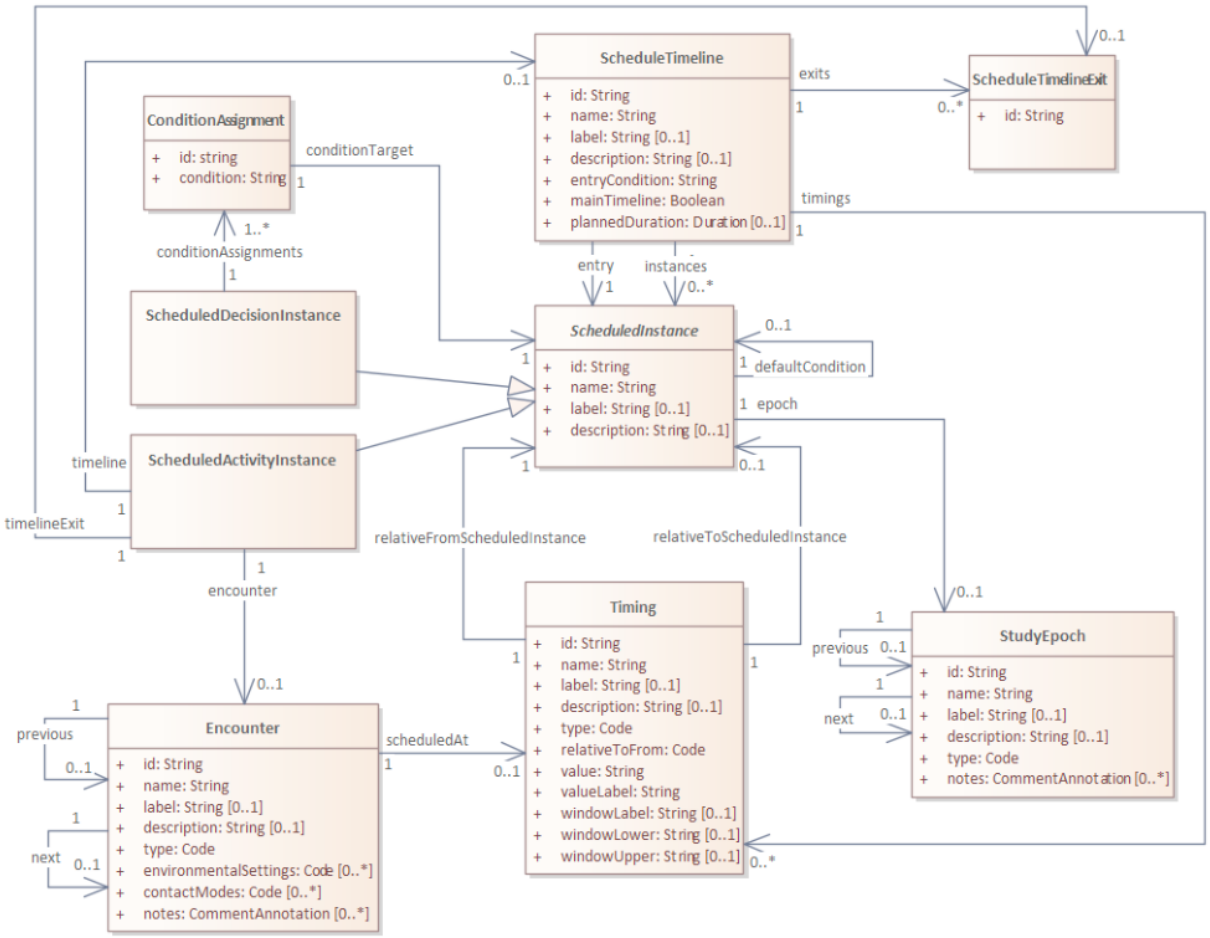

Figure 4 USDM Timeline class describes the Schedule of Activities

This model allows for multiple planned timings within an encounter as well as for decision points in the study process. The corresponding information is stored in a timeline as scheduled activity instances and scheduled decision instances, respectively. Both inherit all attributes and relationships from the ScheduledInstance class (indicated by the closed arrows in the UML) and can be linked to the corresponding study epoch.

The Timing class includes all timing information with details on time between instances and corresponding windowing. One or more scheduled activity instance can be related to a corresponding encounter, which is usually presented as a visit in the schedule of activities.

SYNTHETIC DATA WORKFLOW

This section introduces the end-to-end workflow for generating a synthetic subject journey through the study, and outlines the key steps involved. The subject journey description is then used to generate individual dataframes, and is the key to ensuring consistency in line with the study design.

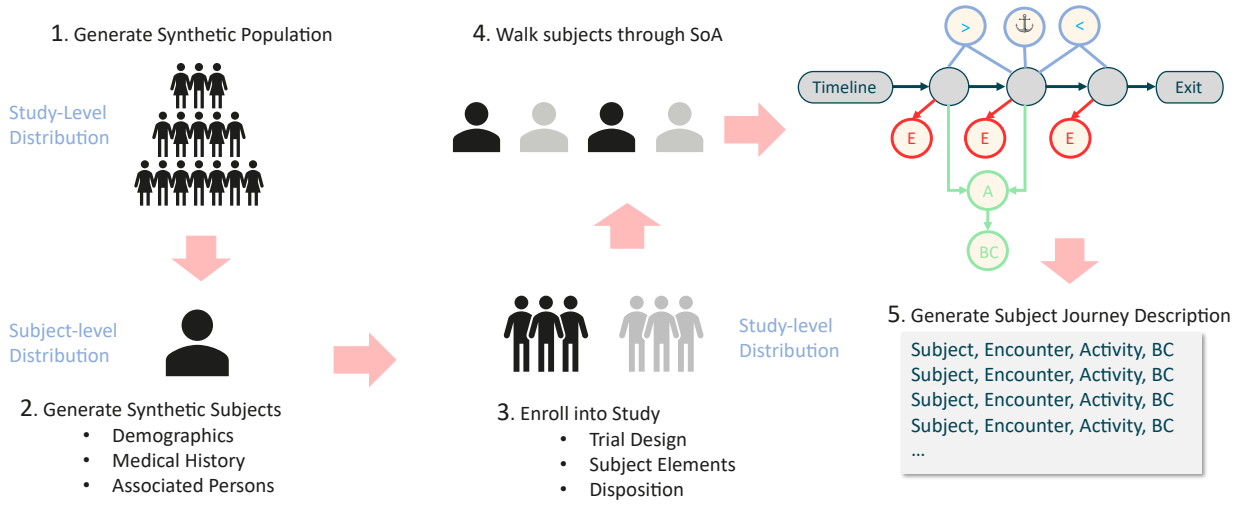

Figure 5 Overview of generation of a subject journey through the study

The key steps are:

1. **Define Synthetic Population:**
   Specify the target population's characteristics, such as demographics and stratification, based on the protocol. This ensures the synthetic cohort accurately reflects the intended study population and supports meaningful downstream analysis.
2. **Generate Synthetic Subjects:**
   Create individual subject profiles with relevant attributes, including medical history and associated persons. Each subject is uniquely constructed to represent realistic variability within the defined population.
3. **Enroll Subjects into Study:**
   Assign subjects to study arms, elements, and disposition according to the trial design. This step ensures that each subject's participation aligns with protocol-defined pathways and eligibility.
4. **Model Subject Journeys:**
   Simulate each subject's path through the Schedule of Activities, capturing visits, interventions, and assessments. The subject journey is mapped in detail to reflect protocol logic and timing.
5. **Generate Subject Journey Descriptions:**
   Produce an intermediate file that describes each subject's journey through the study, capturing all relevant events and transitions. This structured description serves as the foundation for downstream synthetic data generation at the dataframe level, ensuring consistency and protocol alignment.

FINDINGS AND FUTURE DIRECTIONS

This project is a work-in-progress, with key research, design work and prototyping been done to establish feasibility, key findings and next steps include:

- USDM-driven synthetic data generation can be used to generate more realistic test data that is consistent with the study design.
- Separating study modelling from data generation increases flexibility and reusability. The User Journey description format is a key interface and may benefit from an open-source approach.
- Key challenges include managing unstructured text, adding AI/ML to generation, and improving user interfaces for simulation control.

CONTACT INFORMATION

Contact the author at:
Author Name: Stuart Malcolm
Company: Veramed
Email: stuart.malcolm@veramed.co.uk
Website: www.veramed.co.uk
LinkedIn: www.linkedin.com/in/stuart-malcolm/

Brand and product names are trademarks of their respective companies.