Generating Synthetic **Clinical Trial** Data: System Requirements.

This document is a compilation of requirements for software to programmatically generate Synthetic Clinical Trial data for both research as well as review (regulatory) purposes.

This work is sponsored by PHUSE. All documentation, software, and sample output data files created under this project is available for non-commercial use only under an Open Source license (GPLv3) at <https://github.com/phuse-org/PODR/tree/master/sample_code>

# Business Objective

To provide high quality clinical trial data at scale. With no legal restrictions, and no privacy issues. Through the programmatic generation of scientifically-accurate, realistic, yet completely random Clinical Trial data that meets 02 Use Cases:

\* Use Case 01: Synthetic Clinical Trial Data for Software Testing

\* Use Case 02: Synthetic Clinical Trial Data for Regulatory Review

## Project Structure

There are 04 components to this project:

| **Component** | **Description** | **Responsible Parties** |
| --- | --- | --- |
| Inputs Definition | Subject Matter Experts (“SMEs”) define the specific elements of interest when designing a clinical trial. | PHUSE members, FDA |
| Output Files | Text files that mimic the results of a clinical trial, based on the Inputs Definition entered by user. Files match industry standards specifications of CDISC initially; FIHR, HL7 and OMOP later on. | NIHPO |
| User Interface | Allows user to enter Input Definition using either a Microsoft Excel template and/or a web interface. | NIHPO |
| Processing Engine | Software platform that receives Input Definitions through a User Interface and programmatically generates Output Files. | NIHPO |

# Contact Us

Jose C. Lacal, CTO

NIHPO, Inc.

+1 (561) 777-2577

Jose.Lacal@NIHPO.com

# Definitions

These are the definitions of concepts and terms used in this document.

## Synthetic Data

Synthetic data is randomly generated based on pre-defined rules that mimic the statistical properties of real data. Synthetic data is designed to act as a proxy for real data.[[1]](#footnote-2)

# The Business Case for Synthetic **Clinical Trial** Data[[2]](#footnote-3)

Having early, easy access to comprehensive clinical trial results data is necessary for the advancement of health care in general, and drug discovery and approval in particular. However, both sponsors as well as regulators have struggled to easily acquire and manage the required amounts of test data to help accelerate research and review activities. "Synthetic" clinical trial data is designed to help fill the gaps to promote drug discovery and approval.

Synthetic clinical trial records are data sets that contain the trial results of health records of realistic -but not real- patients. Users of this software will be able to control how comprehensive they make the records.

Barriers to accessing real clinical trial data can be immense, but that trial data is vital to conduct research, plan public health measures, and develop effective health IT applications. Still, patient data can be costly to access, and researchers often need a lot of it to build accurate models. There are also significant privacy concerns surrounding patient data; waiting for approval from an independent review board slows down research and development, and there have been cases of anonymized patient records being re-identified.

Enter synthetic data. It is typically free, immediately accessible in large quantities, and no one worries about the privacy of fake patients. Moreover, there are some things that can be done with synthetic data that just aren't possible with actual health records. For example, the data can be modified to fit certain demographic populations or include specific pieces of health information.

That said, synthetic data is not without its limitations. Jason Walonoski, co-creator of Synthea, has said that synthetic data should not be used for clinical discovery, and that researchers should always go back to real data to verify their results. A recent study also concluded that while Synthea does a good job of modeling "average" health care encounters, it has trouble accounting for variations in care.

Who is using synthetic health data?

Consider how your health system can benefit from synthetic health records. Possibilities include:

\* Arming your clinicians with a safe and robust new source of data for early to middle stages of research;

\* Giving developers a new resource for creating systems to improve hospital functions; and

\* Deploying it as a secure tool for teaching medical students and health IT professionals using realistic patient records.

# Use Case 01: Synthetic Clinical Trial Data for Software Testing

Research purposes

A.) Development project inside clinical workflows.

+ Driven by data availability.

+ Once trial ends there is an urgency to process collected data.

+ Large amounts of data (clear and complete) are needed to \*validate\* all \*clinical scenarios\*.

+ “Wait for data” and then rush to “close up study.”

+ Starting implementation of development of analysis late, creates urgent tasks.

+ Teams start coding early but have no test data to play with.

=> Better shape at end of study if coding and testing start early.

B.) Increasing value add.

+ In the last 05 years, both sponsors as well as consultants / CROs, much more attention paid to software application development. Taking a “software approach.”

+ Offering beyond basic services.

+ Moving from decades of single-use scripting to full-blown software development.

+ “Without robust data you can’t have robust testing.”

=> Synthetic Data must have familiar anomalies that real clinical data has: missing values, out of range entries. Lead to stress-testing software, assumptions.

When creating Synth Data: include capability to define “data anomalies.”

How can we do this?

At the time when there was already urgency to close the study.

>> Mimic “Plan Study Data.”

Default current behavior: look for previous study’s data, make minor changes.

Outcome of this approach always exposes itself when real data arrives.

Tremendous amount of effort to achieve a more reliable model.

Recurring challenge.

“We just wait for the arrival of suitable data.”

Software development teams:

+ Using tools that require data much earlier.

+ They need to start coding sooner, they need “realistic data” and to test boundaries, rules.

PHUSE: Standards Analysis and Code Sharing working group.

## System Parameters

+ Online

+ Simple UI

+ Input “population characteristics” based on the study’s inclusion and exclusion criteria

+ Not proprietary

## Output Formats

CDISC-compliant output files:

+ SEND (simple, less domains)

+ ADaM

+ SDTM

Most important files:

+ DM (ADSL)

+ AE (ADAE)

Domains:

+ Disposition: with different trial milestones

+ EC/EX: with varying dose forms and scenarios

## Deployment Models

“Paid and Maintained” service.

Desktop version as well => privacy and security considerations.

Container / VM instead?

Features:

Provide general specs on study design that can be modeled in relationships across variables / visits.

Example: cross-over combinations treatment arms.

Lots of thoughts contributed.

Best recommendations for examining clinical trial data.

=> Use as requirements to build software for industry to use.

= = Test Data Factory = =

+ Trial Design Matrix in an Excel file.

How to model variables?

Synthetic Data becomes challenging.

SEND pre-clinical data.

|=> Synthetic Data should hold up to some “credibility tests.”

# Use Case 02: Synthetic Clinical Trial Data for Regulatory Review

Review purposes

<FDA feedback required>

## Related Projects

These are existing projects with similar goals to those listed in this document.

## Synthea

<https://synthetichealth.github.io/synthea/>

## Relevant Research and Publications

This is a brief list of papers relevant to this topic.

Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record

<https://academic.oup.com/jamia/article/25/3/230/4098271>

Data-driven approach for creating synthetic electronic medical records

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2972239/>

Synthetic Event Time Series Health Data Generation

<https://arxiv.org/pdf/1911.06411v1.pdf>

The Ultimate Guide to Synthetic Data in 2020

<https://research.aimultiple.com/synthetic-data/>

1. Paraphrased from http://www.ehealthinformation.ca/faq/ [↑](#footnote-ref-2)
2. Paraphrased from https://www.advisory.com/research/health-care-it-advisor/it-forefront/2019/11/synthetic-health-data [↑](#footnote-ref-3)