Generating Synthetic Clinical Trial Data: System Requirements.

We intend to build a software platform 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 are available for non-commercial use under an Open Source license (GPLv3)[[1]](#footnote-1).

# Business Objective

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

\* Use Case 01: Synthetic Clinical Trial Data for Process Validation

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

## Project Structure

There are 04 arms to this project:

| Arm | Description | Responsible Parties |
| --- | --- | --- |
| Inputs Definition | Subject Matter Experts (“SMEs”) define the specific parameters 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.  Output Files will match industry standards specifications. CDISC initially. FIHR, HL7 and/or OMOP later on. | NIHPO |
| Online User Interface | Web-based interface allows user to enter Input Definition and define Output Files. | VCA Plus[[2]](#footnote-2) |
| Processing Engine | Software platform that receives Input Definitions and then programmatically generates the requested 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 Clinical Trial Data

Synthetic clinical trial data is randomly generated based on pre-defined rules that mimic the statistical properties of real clinical trial data. Synthetic clinical trial data is designed to act as a proxy for real data.[[3]](#footnote-3)

# Development Environment

You need to configure your local computer as follows:

\* Install a suitable text editor. We strongly recommend Sublime[[4]](#footnote-4)

\* Install Python 3.x

\* Download latest Python script[[5]](#footnote-5)

\* Download SQLite3 file[[6]](#footnote-6)

You should be comfortable using the command line interface (“CLI”) of your chosen operating system.

## Web-based Environment

An Online User Interface will provide an easy-to-use mechanism for users to obtain Output Files without the need to install a local Development Environment.

The picture below gives a high-level overview on the proposed structure showing the following components of the solution.

* The user interface accessible through a standard web browser.
* The script-based wrapper that executes on the Vicos™ platform and translates the user input from the user interface into the parameters required by the scripts that generate the Synthetic Clinical Trial data.
* The Synthetic Clinical Trial Data scripts.
* The cloud-hosted Vicos™ platform, developed and provided by VCA-Plus, Inc with script repositories and the script execution engines.

  
Figure: Overview of Web-based Environment

When using the web-based version of the Synthetic Clinical Trial Data generator, there will be no need to install any programming environment or knowledge of programming.

# Inputs Definition

These are the initial parameters user can define to customize the Output Files.

| Input Parameter | Description |
| --- | --- |
| Number of Subjects | Desired number of Synthetic Subjects to generate data for. |
| Date of enrollment start | Date (in the past) when enrollment started. |
| Date files are generated | Date used to generate Output Files. This date is used for estimating visit and death dates. |
| Percentage Female / Male | Ratio of female/ male in the distribution of subjects. |
| Minimum, maximum ages | Range of inclusion ages. |
| Race splits | The available choices are: \* AMERICAN INDIAN OR ALASKA NATIVE \* ASIAN \* BLACK OR AFRICAN AMERICAN \* NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER \* WHITE \* NOT REPORTED \* UNKNOWN |
| Ethnicity split | The available choices are: \* Hispanic \* Non-Hispanic |
| Arm identifiers | Assign Subjects by percentage of total enrollment to each arm. |
| Group identifiers | Assign Subjects by percentage of total enrollment to each group. |
| Site identifiers | Assign Subjects by percentage of total enrollment to each site. |
| Investigator identifiers | Assign Subjects by percentage of total enrollment to each investigator. |
| Percentage deaths | Percentage of Subjects that will die during the trial. |
| Causes of death | Distribution of causes of death. |
| Drop off rates | Percentage of subjects that do not finish each phase of the trial. |
| Adverse event rate | Percentage of subjects that experience at least 01 adverse event. |
| Country enrollment | Assign Subjects by percentage of total enrollment to each country. |

Please insert additional parameters of interest.

## Structure of Trial

User can define the specific structure of the trial. As follows:

\* Define number of Visits.

For each Visit, define number of Analysis.

For each Analysis, define number of Parameters.

For each Parameter: define Lower value, Upper value, and Fuzz factor.

Here is an example of how the Processing Engine represents a trial structure as an array:

{"visits": [

{"visit\_id" : "Visit\_01", "visit\_name" : "Visit 01 Name", "days\_after\_enrollment": 5, "participation\_rate": 95, "analysis\_list" : [

{"analysis\_id": "Analysis\_01", "analysis\_name": "Visit 01 Name - Analysis 01 Name", "parameter\_list" : [

{"parameter\_id": "Parameter\_01", "parameter\_name": "Visit 01 Name - Analysis 01 - Parameter Name 01", "days\_delay": 12, "value\_list": [

{"Lower limit": 12},

{"Upper limit": 45},

{"Fuzz factor": 0.35},

]},

{"parameter\_id": "Parameter\_02", "parameter\_name": "Visit 01 Name - Analysis 01 - Parameter Name 02", "days\_delay": 1, "value\_list": [

{"Lower limit": 120},

{"Upper limit": 450},

{"Fuzz factor": 0.15},

]},

{"parameter\_id": "Parameter\_03", "parameter\_name": "Visit 01 Name - Analysis 01 - Parameter Name 03", "days\_delay": 6, "value\_list": [

{"Lower limit": 0.97},

{"Upper limit": 2.67},

{"Fuzz factor": 0.47},

]

}]

},

},

]}

# Open Items

These are items that need to be addressed

## Credibility Tests

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

## Data Anomalies

Synthetic Data must have familiar anomalies that real clinical data has: missing values, out of range entries. Lead to stress-testing software, assumptions. Include capability to define “data anomalies.”

# Use Case 01: Synthetic Clinical Trial Data for Process Validation

|  |  |
| --- | --- |
| Summary | User will be able to define a full Clinical Trial, fine-tuning each Input Parameter, and can specify the desired Output Files. |
| Actors | This Use Case is specifically focused on sponsors and CROs. Inside these organizations, the intended actors are:  \* Project and Study Biostatisticians  \* Project and Study Statistical Programmers  \* Software development teams  \* Quality control personnel |
| Pre-conditions | \* Set up the Development Environment as described above. |
| Description | \* Actor describes the “Plan Study Data” using the Inputs Definition described above.  \* Actor inputs population characteristics based on the study’s inclusion and exclusion criteria  \* Provide general specs on study design that can be modeled in relationships across variables / visits. For example: cross-over combinations treatment arms. |
| Exceptions | \* Software will alert Actor is any required parameters are missing.  \* Software will display useful and clear exceptions and error messages (such as unable to access a directory). Messages will include suggestions for user to correct exceptions / errors. |
| Post-conditions | \* Actor will receive a series of Output Files in the format specified and that reflect the desired characteristics of the Subjects in the Clinical Trial. |

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

|  |  |
| --- | --- |
| Summary | User will be able to define a full Clinical Trial, fine-tuning each Input Parameter, and can specify the desired Output Files. |
| Actors | This Use Case is specifically focused on regulatory agencies. Inside these organizations, the intended actors are:  \* <..> |
| Pre-conditions | \* Set up the Development Environment as described above. |
| Description | \* Actor describes the expected data using the Inputs Definition described above. |
| Exceptions | \* Software will alert Actor is any required parameters are missing.  \* Software will display useful and clear exceptions and error messages (such as unable to access a directory). Messages will include suggestions for user to correct exceptions / errors. |
| Post-conditions | \* Actor will receive a series of Output Files in the format specified and that reflect the desired characteristics of the Subjects in the Clinical Trial. |

<FDA feedback required>

For example:

a.) Provide guidance on how to best implement the “Study Data Standards Resources”

[https://www.fda.gov/industry/fda-resources-data-standards/study-data-standards-resources](Synthetic_Clinical_Data_Requirements_02.docx)

[https://www.fda.gov/industry/study-data-standards-resources/study-data-submission-cder-and-cber](Synthetic_Clinical_Data_Requirements_02.docx)

b.) Define types (and format) of reports the agency expects to receive from sponsors.

## 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/](Synthetic_Clinical_Data_Requirements_02.docx)

## Books

Practical Synthetic Data Generation

by Khaled El Emam, Lucy Mosquera, Richard Hoptroff

Released May 2020

Publisher(s): O'Reilly Media, Inc.

ISBN: 9781492072744

[https://www.oreilly.com/library/view/practical-synthetic-data/9781492072737/](Synthetic_Clinical_Data_Requirements_02.docx)

1. https://github.com/phuse-org/PODR/tree/master/sample\_code [↑](#footnote-ref-1)
2. http://www.vca-plus.com [↑](#footnote-ref-2)
3. Paraphrased from http://www.ehealthinformation.ca/faq/ [↑](#footnote-ref-3)
4. http://www.sublimetext.com/ [↑](#footnote-ref-4)
5. https://github.com/phuse-org/PODR/tree/master/sample\_code [↑](#footnote-ref-5)
6. https://github.com/phuse-org/PODR/blob/master/sample\_code/Synthetic\_Health\_Data\_NIHPO.sqlite3 [↑](#footnote-ref-6)