PharmaSUG 2021 - Paper AD-162

Generating Synthetic CDISC Clinical Trial Data.

José C. Lacal; NIHPO, Inc.

ABSTRACT

This paper presents a Python-based platform that generates Synthetic CDISC Clinical Trial Data at scale. The platform programmatically generates realistic synthetic persons (SynthPerson™) that have a complete synthetic Personal Health Record (SynthPHR™). Each SynthPerson™ lives in a series of geographically-accurate synthetic cities (SynthCity™). Groups of SynthPerson™ are then randomly enrolled in a synthetic Clinical Trial (SynthTrial™). Platform users define the parameters of each synthetic clinical trial and the platform generates all CDISC SDTM domains for the desired number of subjects.

INTRODUCTION

This paper describes the structure and functionality of a Python-based platform that generates synthetic health data at scale. One of the output types the platform generates are standard-compliant CDISC SDTM files in SAS Transport File Format (XPORT).

First, the platform allows users to generate very realistic yet fake synthetic individuals ("SynthPerson™"). Each SynthPerson™ receives a complete individual demographic profile including date of birth, gender, and places of birth and residence (SynthCity™), among other user-defined variables.

Second, the platform generates a full Personal Health Record (SynthPHR™) for each SynthPerson™. Each SynthPHR™ includes both Real-World Data from government agencies (EMA, FDA) as well as random values assigned from controlled terminologies (LOINC, SNOMED-CT). The goal of each SynthPHR™ is to provide a life-long, realistic, comprehensive medical history for each SynthPerson™.

Third, user can define the parameters for a synthetic clinical trial (SynthTrial™), virtually enrolling the synthetic subjects defined previously. The platform generates synthetic results for the clinical trial, based on the number of epochs, visits, arms, etc. defined by user. User can play with the clinical trial parameters and can quickly generate output files for different scenarios of the same trial.

Finally, the platform generates different types of output files: CSV, JSON, SAS (xport), and SQLite.

The data generated by this platform is not intended to replace "real" healthcare data. Rather, this platform wants to encourage and facilitate the use of synthetic health data as a temporary placeholder for real data. We believe synthetic health data can be useful to accelerate and shorten all test, QA, and end-to-end system validation in life science applications. The platform's initial focus is to provide synthetic health data across the lifecycle of a clinical trial.

It is worth mentioning that the platform explicitly, purposefully generates fully random data where, for example, a SynthPerson™ with gender equal to "Male" may be assigned a condition of "pregnancy". This synthetic data is designed to test assumptions and rules built into software used with clinical trial data.

The paper is divided in the following sections:

- Market Need
- Creating a Synthetic Person (SynthPerson™)
- Building a Synthetic Personal Health Record (SynthPHR™)
- Building a Synthetic Clinical Trial (SynthTrial™)
- Future Work and Extensions
- Tools
- References

We close the paper with Conclusions and Acknowledgements.

By the end of the paper the reader will gain an understanding of the structure of the platform's architecture, written in Python3. And reader will be able to create synthetic subjects, synthetic cohorts, and synthetic clinical trials, for free, using the platform's source code available in GitHub.

MARKET NEED

The author identified the need for a tool that generates synthetic health data at scale through his work with the PHUSE Test Data Factory working group.

The motivation behind this platform is to allow all staff levels at an organization participating in clinical trials to easily create and use synthetic data with no restrictions. It is our hope that this platform will accelerate the development and approval of both new medical devices and pharmaceutical products.

THE PROBLEM WE SOLVE

We became aware that many players in the life sciences industry are unable to easily access realistic, unencumbered health data at scale before the real clinical trial data becomes available.

This non-real (yet realistic) synthetic health data can be of assistance during software development and testing. And to evaluate corner cases.

Early adopters report that synthetic health data can potentially shorten Quality Assurance cycles, and also provide a crucial tool for end-to-end System Validation.

Finally, and very importantly, this synthetic health data is available with no copyright, legal, privacy, or regulatory blocks.

OTHER APPROACHES

There are several mechanisms life science organizations rely on to obtain synthetic health data:

- Anonymized real health data. There are multiple tools that anonymize health data.
- Models extracted from real health data. Companies such as Replica Analytics take real health data and through proprietary mechanism the company generates a "model" based on the real data. Such model is then used to generate synthetic data.

In terms of synthetic data, Table 1. Synthetic Health Data generators summarizes a few available synthetic health data generators.

Name	URL	Comments
EMRBOTS.ORG	http://www.emrbots.org	Pre-generated EMR records.
JamesMarcogliese/ Patient-Generator	https://github.com/JamesMarcogliese/ Patient-Generator	Archived by owner.
Synthea	https://synthetichealth.github.io/synthea/	Gold standard.
The Random Patient Generator	https://randompatientgenerator.netlify.app	Single record at a time.

Table 1. Synthetic Health Data generators

We believe our approach is more flexible, scalable, and easier to implement than existing synthetic data generators. Secondly, the synthetic health data this platform generates has absolutely no risk of ever becoming "de-anonymized" as there are no "real" people behind the data. Each anonymization or model-generation tool has an implied de-anonymization risk factor. And the platform is free, under the Affero GPL (AGPL) Open Source license.

CREATING A SYNTHETIC PERSON (SYNTHPERSON™)

The first stage in the platform allows users to generate synthetic persons (SynthPerson™) with a complete demographic profile for each SynthPerson™. The platform can generate large numbers of records with user-defined parameters such as:

- age range (minimum and maximum ages)
- gender distribution (percentages of Female / Male records)
- race distribution (Black, Hispanic, White, other)
- geographical distribution by country and province / state (including place of birth, place of residence) defined using real locations that includes Latitude and Longitude for GIS uses (SynthCity™). We use US Locations and World Locations data from the US government.
- realistic, yet synthetic full name

Think of the platform as a tool that helps users to generate "synthetic cohorts" on demand.

At the end of this stage the user will have files with hundreds (or hundreds of thousands) of synthetic individuals. The synthetic person output is available in different file formats: CSV, JSON, and/or SQLite.

BUILDING A SYNTHETIC PERSONAL HEALTH RECORD (SYNTHPHR™)

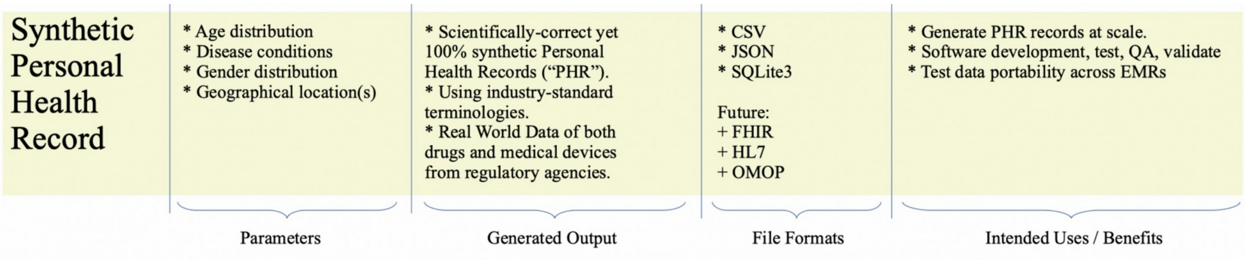
Once a user has built a cohort of synthetic subjects (SynthPerson™), the second stage in the platform builds a full (synthetic) Personal Health Record (SynthPHR™) for each subject.

When generating the PHR records, the platform utilizes scientifically-valid Controlled Terminologies (including SNOMED-CT CORE, LOINC, and CDISC's own Controlled Terminology) as well as Real-World Data sources (including from the FDA, the European Medicines Agency, and the Center for Medicare and Medicaid Services' ("CMS") National Provider Identifier.

There are versions of many controlled terminologies for languages other than English that can be easily plugged into the platform. The platform is then able to generate PHRs in multiple languages.

Finally, the platform outputs the PHR records in multiple formats, including: CSV, JSON, and SQLite. Additional output formats (FHIR, HL7, OMOP) can be easily added to the platform.

Figure 1: Synthetic Personal Record (SynthPHR™) overview below illustrates the core capabilities of the synthetic PHR generation stage.



	Parameters	Generated Output	File Formats	Intended Uses / Benefits
Synthetic Personal Health Record	* Age distribution * Disease conditions * Gender distribution * Geographical location(s)	* Scientifically-correct yet 100% synthetic Personal Health Records ("PHR"). * Using industry-standard terminologies. * Real World Data of both drugs and medical devices from regulatory agencies.	* CSV * JSON * SQLite3 Future: + FHIR + HL7 + OMOP	* Generate PHR records at scale. * Software development, test, QA, validate * Test data portability across EMRs

Figure 1: Synthetic Personal Record (SynthPHR™) overview

The software platform is divided into "Functional Blocks" for ease of development and re-use.

We'll now analyze each functional block's capabilities.

Table 2 below summarizes the functional blocks used to generate Synthetic Personal Health Records.

Functional Block	Purpose: Randomly assigns to each subject
Demographics	Generates records of each Synthetic Person
Conditions	SNOMED-CT CORE-defined medical conditions
Devices	FDA-approved medical devices
Drugs	EMA- or FDA-approved pharmaceutical products
Lab Results	LOINC-defined lab results
Procedures	SNOMED-CT CORE-defined procedures
Providers	Current Medicare-registered healthcare providers
Vitals	Vital signs

Table 2. Functional Blocks used to generate Personal Health Records

It is important to note that each functional block listed above is available through an API endpoint. This software architecture allows each functional block to be independently re-used for multiple purposes, as we'll cover in 03. Re-use Functional Blocks below.

DEMOGRAPHICS

This functional block generates the synthetic person record. The Patient_ID field is used as the Foreign Key across the synthetic PHR tables.

Display 1: SynthPHR™ - Demographics below shows the Demographics section of a SynthPHR™ SQLite file.

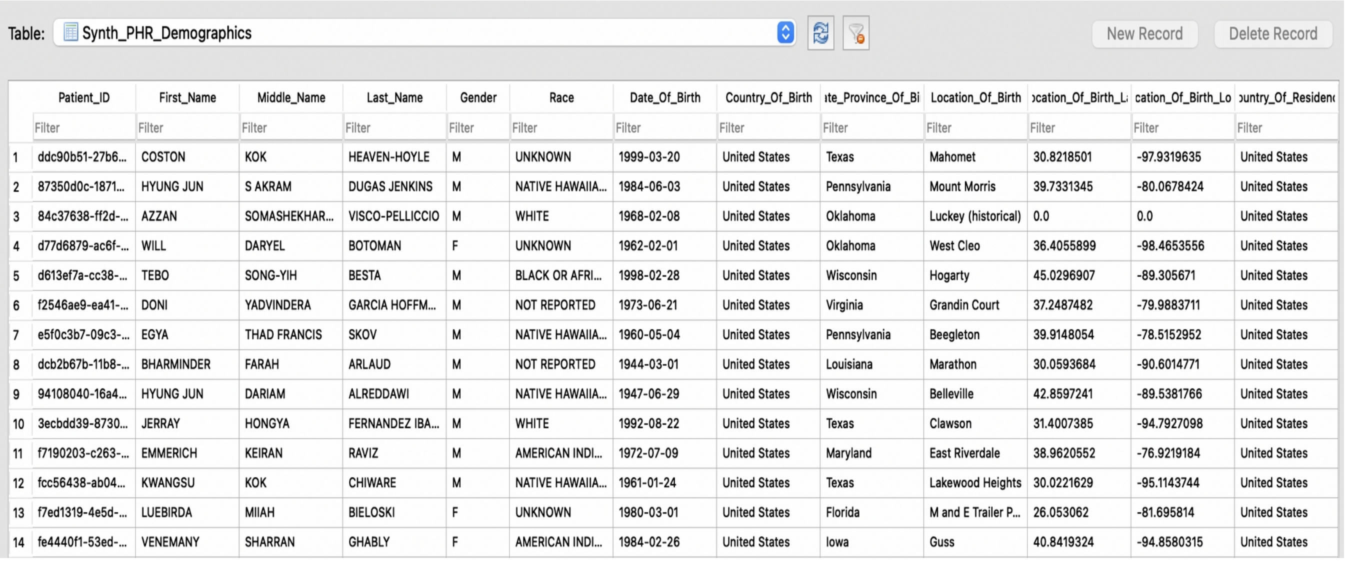


Table: Synth_PHR_Demographics

	Patient_ID	First_Name	Middle_Name	Last_Name	Gender	Race	Date_Of_Birth	Country_Of_Birth	ıte_Province_Of_Bi	Location_Of_Birth	ɔcation_Of_Birth_Lː	cation_Of_Birth_Lo	ɔuntry_Of_Residenc
	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	ddc90b51-27b6...	COSTON	KOK	HEAVEN-HOYLE	M	UNKNOWN	1999-03-20	United States	Texas	Mahomet	30.8218501	-97.9319635	United States
2	87350d0c-1871...	HYUNG JUN	S AKRAM	DUGAS JENKINS	M	NATIVE HAWAIIA...	1984-06-03	United States	Pennsylvania	Mount Morris	39.7331345	-80.0678424	United States
3	84c37638-ff2d-...	AZZAN	SOMASHEKHAR...	VISCO-PELLICCIO	M	WHITE	1968-02-08	United States	Oklahoma	Luckey (historical)	0.0	0.0	United States
4	d77d6879-ac6f-...	WILL	DARYEL	BOTOMAN	F	UNKNOWN	1962-02-01	United States	Oklahoma	West Cleo	36.4055899	-98.4653556	United States
5	d613ef7a-cc38-...	TEBO	SONG-YIH	BESTA	M	BLACK OR AFRI...	1998-02-28	United States	Wisconsin	Hogarty	45.0296907	-89.305671	United States
6	f2546ae9-ea41-...	DONI	YADVINDERA	GARCIA HOFFM...	M	NOT REPORTED	1973-06-21	United States	Virginia	Grandin Court	37.2487482	-79.9883711	United States
7	e5f0c3b7-09c3-...	EGYA	THAD FRANCIS	SKOV	M	NATIVE HAWAIIA...	1960-05-04	United States	Pennsylvania	Beegleton	39.9148054	-78.5152952	United States
8	dcb2b67b-11b8-...	BHARMINDER	FARAH	ARLAUD	M	NOT REPORTED	1944-03-01	United States	Louisiana	Marathon	30.0593684	-90.6014771	United States
9	94108040-16a4...	HYUNG JUN	DARIAM	ALREDDAWI	M	NATIVE HAWAIIA...	1947-06-29	United States	Wisconsin	Belleville	42.8597241	-89.5381766	United States
10	3ecbdd39-8730...	JERRAY	HONGYA	FERNANDEZ IBA...	M	WHITE	1992-08-22	United States	Texas	Clawson	31.4007385	-94.7927098	United States
11	f7190203-c263-...	EMMERICH	KEIRAN	RAVIZ	M	AMERICAN INDI...	1972-07-09	United States	Maryland	East Riverdale	38.9620552	-76.9219184	United States
12	fcc56438-ab04...	KWANGSU	KOK	CHIWARE	M	NATIVE HAWAIIA...	1961-01-24	United States	Texas	Lakewood Heights	30.0221629	-95.1143744	United States
13	f7ed1319-4e5d-...	LUEBIRDA	MIIAH	BIELOSKI	F	UNKNOWN	1980-03-01	United States	Florida	M and E Trailer P...	26.053062	-81.695814	United States
14	fe4440f1-53ed-...	VENEMANY	SHARRAN	GHABLY	F	AMERICAN INDI...	1984-02-26	United States	Iowa	Guss	40.8419324	-94.8580315	United States

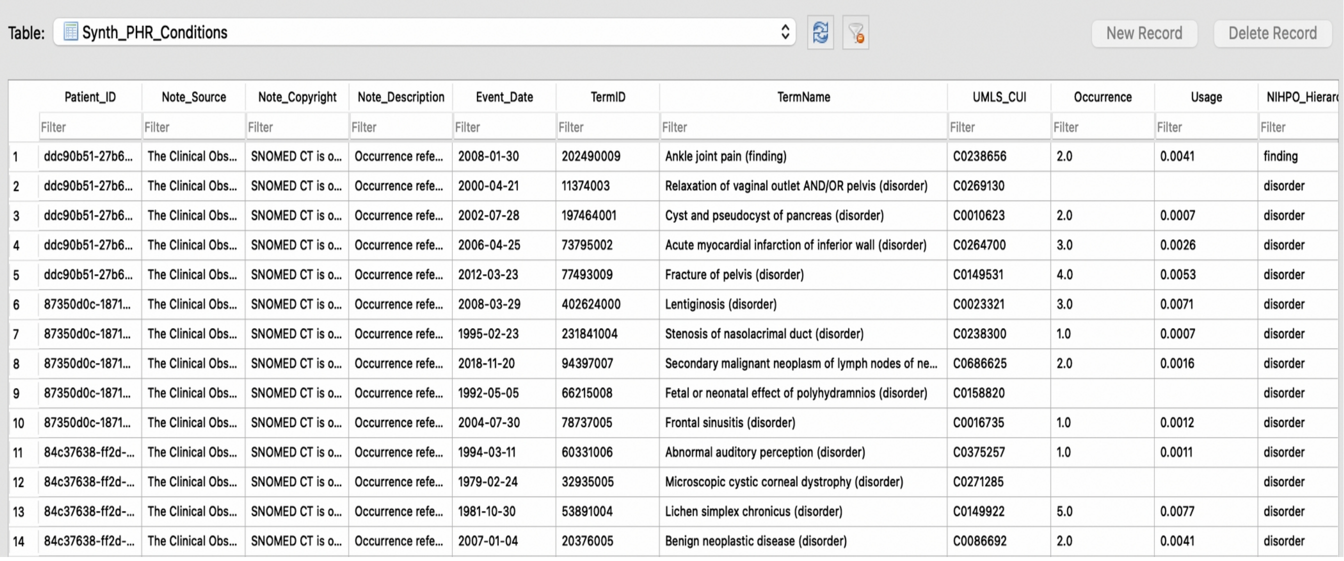
Display 1: SynthPHR™ - Demographics

CONDITIONS

This functional block uses SNOMED-CT CORE, and its 5,316 Findings, to assign random medical conditions to each synthetic subject. Each record has a random event date assigned to the condition.

SNOMED-CT CORE includes fields for "Occurrence" (number of institutions having this concept on their problem list) and "Usage" (the average usage percentage among institutions). The platform uses both Occurrence and Usage to generate random records with the proper statistical frequency.

Display 2: SynthPHR™ - Conditions below shows the Conditions section of a SynthPHR™ SQLite file.



	Patient_ID	Note_Source	Note_Copyright	Note_Description	Event_Date	TermID	TermName	UMLS_CUI	Occurrence	Usage	NIHPO_Hierar
	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	ddc90b51-27b6...	The Clinical Obs...	SNOMED CT is o...	Occurrence refe...	2008-01-30	202490009	Ankle joint pain (finding)	C0238656	2.0	0.0041	finding
2	ddc90b51-27b6...	The Clinical Obs...	SNOMED CT is o...	Occurrence refe...	2000-04-21	11374003	Relaxation of vaginal outlet AND/OR pelvis (disorder)	C0269130			disorder
3	ddc90b51-27b6...	The Clinical Obs...	SNOMED CT is o...	Occurrence refe...	2002-07-28	197464001	Cyst and pseudocyst of pancreas (disorder)	C0010623	2.0	0.0007	disorder
4	ddc90b51-27b6...	The Clinical Obs...	SNOMED CT is o...	Occurrence refe...	2006-04-25	73795002	Acute myocardial infarction of inferior wall (disorder)	C0264700	3.0	0.0026	disorder
5	ddc90b51-27b6...	The Clinical Obs...	SNOMED CT is o...	Occurrence refe...	2012-03-23	77493009	Fracture of pelvis (disorder)	C0149531	4.0	0.0053	disorder
6	87350d0c-1871...	The Clinical Obs...	SNOMED CT is o...	Occurrence refe...	2008-03-29	402624000	Lentiginosis (disorder)	C0023321	3.0	0.0071	disorder
7	87350d0c-1871...	The Clinical Obs...	SNOMED CT is o...	Occurrence refe...	1995-02-23	231841004	Stenosis of nasolacrimal duct (disorder)	C0238300	1.0	0.0007	disorder
8	87350d0c-1871...	The Clinical Obs...	SNOMED CT is o...	Occurrence refe...	2018-11-20	94397007	Secondary malignant neoplasm of lymph nodes of ne...	C0686625	2.0	0.0016	disorder
9	87350d0c-1871...	The Clinical Obs...	SNOMED CT is o...	Occurrence refe...	1992-05-05	66215008	Fetal or neonatal effect of polyhydramnios (disorder)	C0158820			disorder
10	87350d0c-1871...	The Clinical Obs...	SNOMED CT is o...	Occurrence refe...	2004-07-30	78737005	Frontal sinusitis (disorder)	C0016735	1.0	0.0012	disorder
11	84c37638-ff2d-...	The Clinical Obs...	SNOMED CT is o...	Occurrence refe...	1994-03-11	60331006	Abnormal auditory perception (disorder)	C0375257	1.0	0.0011	disorder
12	84c37638-ff2d-...	The Clinical Obs...	SNOMED CT is o...	Occurrence refe...	1979-02-24	32935005	Microscopic cystic corneal dystrophy (disorder)	C0271285			disorder
13	84c37638-ff2d-...	The Clinical Obs...	SNOMED CT is o...	Occurrence refe...	1981-10-30	53891004	Lichen simplex chronicus (disorder)	C0149922	5.0	0.0077	disorder
14	84c37638-ff2d-...	The Clinical Obs...	SNOMED CT is o...	Occurrence refe...	2007-01-04	20376005	Benign neoplastic disease (disorder)	C0086692	2.0	0.0041	disorder

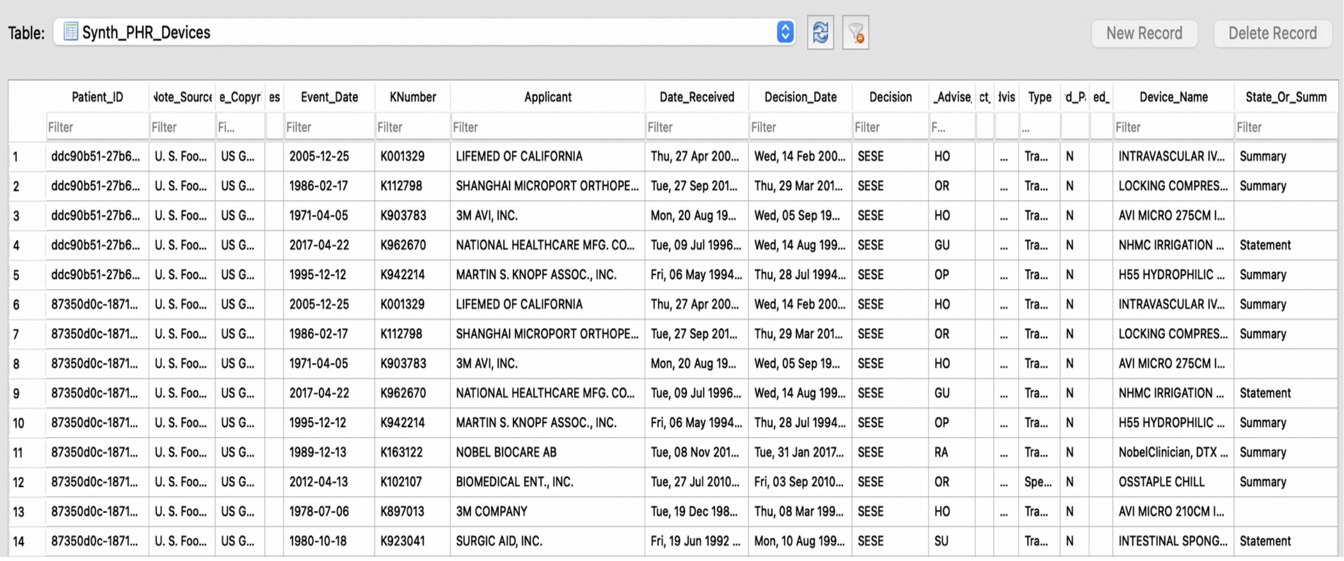
Display 2: SynthPHR™ - Conditions

DEVICES

This functional block takes Real-World Data from the Food and Drug Administration's ("FDA") list of Approved Devices and randomly assigns one or multiple medical devices to each synthetic person's record.

Each device record also receives a random event date when the subject had contact with the assigned medical device.

Display 3: SynthPHR™ - Devices below shows the Devices section of a SynthPHR™ SQLite file.



	Patient_ID	Note_Source	e_Copyr	es	Event_Date	KNumber	Applicant	Date_Received	Decision_Date	Decision	_Advise	ct_	dvis	Type	d_P	ed_	Device_Name	State_Or_Summ
	Filter	Filter	Fi...		Filter	Filter	Filter	Filter	Filter	Filter	F...			...			Filter	Filter
1	ddc90b51-27b6...	U. S. Foo...	US G...		2005-12-25	K001329	LIFEMED OF CALIFORNIA	Thu, 27 Apr 200...	Wed, 14 Feb 200...	SESE	HO		...	Tra...	N		INTRAVASCULAR IV...	Summary
2	ddc90b51-27b6...	U. S. Foo...	US G...		1986-02-17	K112798	SHANGHAI MICROPORT ORTHOPE...	Tue, 27 Sep 201...	Thu, 29 Mar 201...	SESE	OR		...	Tra...	N		LOCKING COMPRES...	Summary
3	ddc90b51-27b6...	U. S. Foo...	US G...		1971-04-05	K903783	3M AVI, INC.	Mon, 20 Aug 19...	Wed, 05 Sep 19...	SESE	HO		...	Tra...	N		AVI MICRO 275CM I...	
4	ddc90b51-27b6...	U. S. Foo...	US G...		2017-04-22	K962670	NATIONAL HEALTHCARE MFG. CO...	Tue, 09 Jul 1996...	Wed, 14 Aug 199...	SESE	GU		...	Tra...	N		NHMC IRRIGATION ...	Statement
5	ddc90b51-27b6...	U. S. Foo...	US G...		1995-12-12	K942214	MARTIN S. KNOPF ASSOC., INC.	Fri, 06 May 1994...	Thu, 28 Jul 1994...	SESE	OP		...	Tra...	N		H55 HYDROPHILIC ...	Summary
6	87350d0c-1871...	U. S. Foo...	US G...		2005-12-25	K001329	LIFEMED OF CALIFORNIA	Thu, 27 Apr 200...	Wed, 14 Feb 200...	SESE	HO		...	Tra...	N		INTRAVASCULAR IV...	Summary
7	87350d0c-1871...	U. S. Foo...	US G...		1986-02-17	K112798	SHANGHAI MICROPORT ORTHOPE...	Tue, 27 Sep 201...	Thu, 29 Mar 201...	SESE	OR		...	Tra...	N		LOCKING COMPRES...	Summary
8	87350d0c-1871...	U. S. Foo...	US G...		1971-04-05	K903783	3M AVI, INC.	Mon, 20 Aug 19...	Wed, 05 Sep 19...	SESE	HO		...	Tra...	N		AVI MICRO 275CM I...	
9	87350d0c-1871...	U. S. Foo...	US G...		2017-04-22	K962670	NATIONAL HEALTHCARE MFG. CO...	Tue, 09 Jul 1996...	Wed, 14 Aug 199...	SESE	GU		...	Tra...	N		NHMC IRRIGATION ...	Statement
10	87350d0c-1871...	U. S. Foo...	US G...		1995-12-12	K942214	MARTIN S. KNOPF ASSOC., INC.	Fri, 06 May 1994...	Thu, 28 Jul 1994...	SESE	OP		...	Tra...	N		H55 HYDROPHILIC ...	Summary
11	87350d0c-1871...	U. S. Foo...	US G...		1989-12-13	K163122	NOBEL BIOCARE AB	Tue, 08 Nov 201...	Tue, 31 Jan 2017...	SESE	RA		...	Tra...	N		NobelClinician, DTX ...	Summary
12	87350d0c-1871...	U. S. Foo...	US G...		2012-04-13	K102107	BIOMEDICAL ENT., INC.	Tue, 27 Jul 2010...	Fri, 03 Sep 2010...	SESE	OR		...	Spe...	N		OSSTAPLE CHILL	Summary
13	87350d0c-1871...	U. S. Foo...	US G...		1978-07-06	K897013	3M COMPANY	Tue, 19 Dec 198...	Thu, 08 Mar 199...	SESE	HO		...	Tra...	N		AVI MICRO 210CM I...	
14	87350d0c-1871...	U. S. Foo...	US G...		1980-10-18	K923041	SURGIC AID, INC.	Fri, 19 Jun 1992 ...	Mon, 10 Aug 199...	SESE	SU			Tra...	N		INTESTINAL SPONG...	Statement

Display 3: SynthPHR™ - Devices

DRUGS

This functional block takes Real-World Data from either the European Medicines Agency ("EMA") or the FDA's Drugs@FDA and randomly assigns pharmaceutical products to each synthetic subject's record.

Display 4: SynthPHR™ - Drugs below shows the Drugs section of a SynthPHR™ SQLite file.

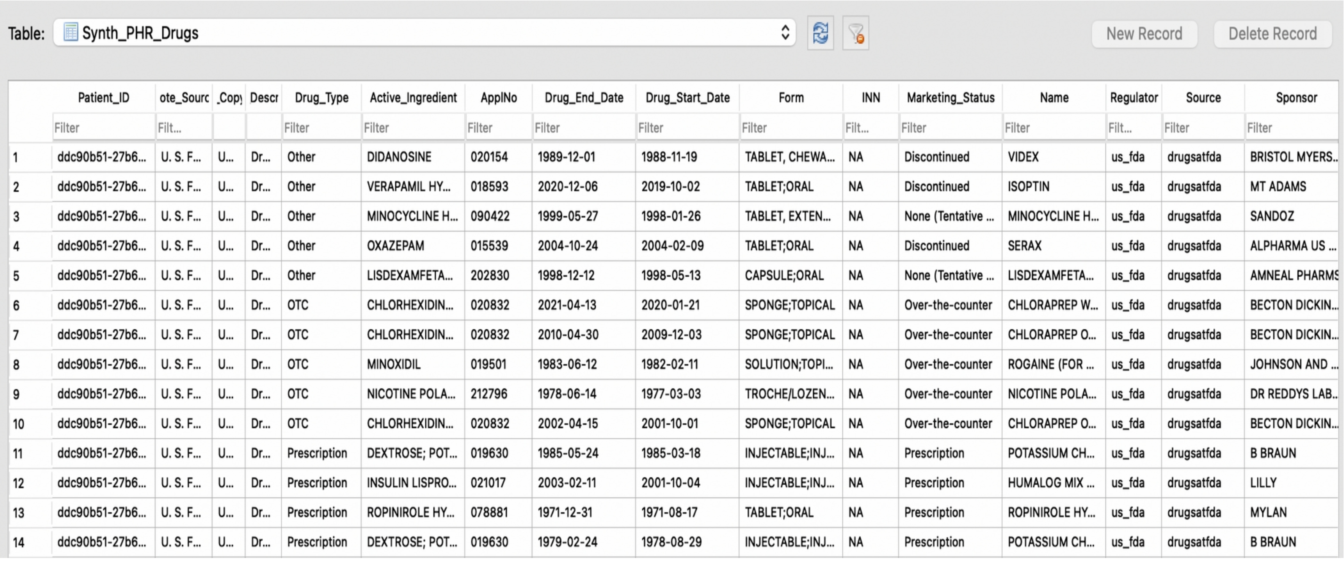


Table: Synth_PHR_Drugs

	Patient_ID	ote_Sourc	_Copy	Descr	Drug_Type	Active_Ingredient	ApplNo	Drug_End_Date	Drug_Start_Date	Form	INN	Marketing_Status	Name	Regulator	Source	Sponsor
	Filter	Filt...			Filter	Filter	Filter	Filter	Filter	Filter	Filt...	Filter	Filter	Filt...	Filter	Filter
1	ddc90b51-27b6...	U. S. F...	U...	Dr...	Other	DIDANOSINE	020154	1989-12-01	1988-11-19	TABLET, CHEWA...	NA	Discontinued	VIDEX	us_fda	drugsatfda	BRISTOL MYERS...
2	ddc90b51-27b6...	U. S. F...	U...	Dr...	Other	VERAPAMIL HY...	018593	2020-12-06	2019-10-02	TABLET;ORAL	NA	Discontinued	ISOPTIN	us_fda	drugsatfda	MT ADAMS
3	ddc90b51-27b6...	U. S. F...	U...	Dr...	Other	MINOCYCLINE H...	090422	1999-05-27	1998-01-26	TABLET, EXTEN...	NA	None (Tentative ...	MINOCYCLINE H...	us_fda	drugsatfda	SANDOZ
4	ddc90b51-27b6...	U. S. F...	U...	Dr...	Other	OXAZEPAM	015539	2004-10-24	2004-02-09	TABLET;ORAL	NA	Discontinued	SERAX	us_fda	drugsatfda	ALPHARMA US ...
5	ddc90b51-27b6...	U. S. F...	U...	Dr...	Other	LISDEXAMFETA...	202830	1998-12-12	1998-05-13	CAPSULE;ORAL	NA	None (Tentative ...	LISDEXAMFETA...	us_fda	drugsatfda	AMNEAL PHARMS
6	ddc90b51-27b6...	U. S. F...	U...	Dr...	OTC	CHLORHEXIDIN...	020832	2021-04-13	2020-01-21	SPONGE;TOPICAL	NA	Over-the-counter	CHLORAPREP W...	us_fda	drugsatfda	BECTON DICKIN...
7	ddc90b51-27b6...	U. S. F...	U...	Dr...	OTC	CHLORHEXIDIN...	020832	2010-04-30	2009-12-03	SPONGE;TOPICAL	NA	Over-the-counter	CHLORAPREP O...	us_fda	drugsatfda	BECTON DICKIN...
8	ddc90b51-27b6...	U. S. F...	U...	Dr...	OTC	MINOXIDIL	019501	1983-06-12	1982-02-11	SOLUTION;TOPI...	NA	Over-the-counter	ROGAINE (FOR ...	us_fda	drugsatfda	JOHNSON AND ...
9	ddc90b51-27b6...	U. S. F...	U...	Dr...	OTC	NICOTINE POLA...	212796	1978-06-14	1977-03-03	TROCHE/LOZEN...	NA	Over-the-counter	NICOTINE POLA...	us_fda	drugsatfda	DR REDDYS LAB...
10	ddc90b51-27b6...	U. S. F...	U...	Dr...	OTC	CHLORHEXIDIN...	020832	2002-04-15	2001-10-01	SPONGE;TOPICAL	NA	Over-the-counter	CHLORAPREP O...	us_fda	drugsatfda	BECTON DICKIN...
11	ddc90b51-27b6...	U. S. F...	U...	Dr...	Prescription	DEXTROSE; POT...	019630	1985-05-24	1985-03-18	INJECTABLE;INJ...	NA	Prescription	POTASSIUM CH...	us_fda	drugsatfda	B BRAUN
12	ddc90b51-27b6...	U. S. F...	U...	Dr...	Prescription	INSULIN LISPRO...	021017	2003-02-11	2001-10-04	INJECTABLE;INJ...	NA	Prescription	HUMALOG MIX ...	us_fda	drugsatfda	LILLY
13	ddc90b51-27b6...	U. S. F...	U...	Dr...	Prescription	ROPINIROLE HY...	078881	1971-12-31	1971-08-17	TABLET;ORAL	NA	Prescription	ROPINIROLE HY...	us_fda	drugsatfda	MYLAN
14	ddc90b51-27b6...	U. S. F...	U...	Dr...	Prescription	DEXTROSE; POT...	019630	1979-02-24	1978-08-29	INJECTABLE;INJ...	NA	Prescription	POTASSIUM CH...	us_fda	drugsatfda	B BRAUN

Display 4: SynthPHR™ - Drugs

LAB RESULTS

This functional block randomly assigns LOINC-defined lab results (including descriptions, measurements, and values) to each subject's record.

Display 5: SynthPHR™ - Lab Results below shows the Lab Results section of a SynthPHR™ SQLite file.

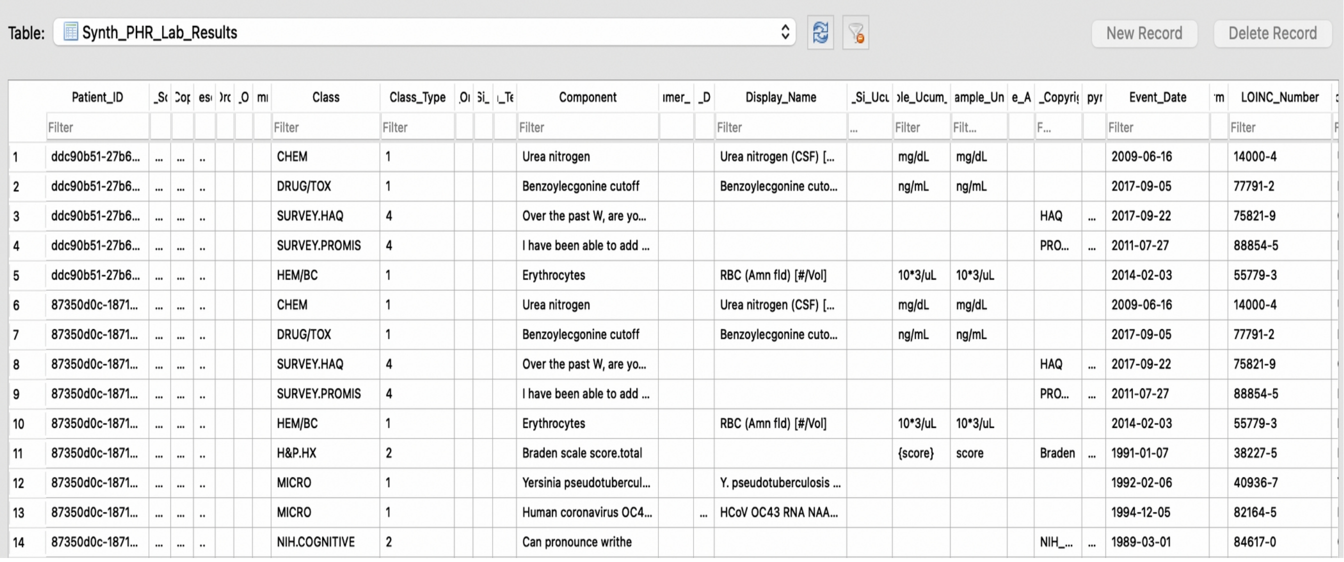


Table: Synth_PHR_Lab_Results

	Patient_ID	_Sc	Cop	es	Drc	_O	mi	Class	Class_Type	_Oı	Si_	_Te	Component	imer_	_D	Display_Name	_Si_Ucu	ble_Ucum_	ample_Un	e_A	_Copyri	pyr	Event_Date	m	LOINC_Number
	Filter							Filter	Filter				Filter			Filter	...	Filter	Filt...		F...		Filter		Filter
1	ddc90b51-27b6...	...	...	..				CHEM	1				Urea nitrogen			Urea nitrogen (CSF) [...		mg/dL	mg/dL				2009-06-16		14000-4
2	ddc90b51-27b6...	...	...	..				DRUG/TOX	1				Benzoylecgonine cutoff			Benzoylecgonine cuto...		ng/mL	ng/mL				2017-09-05		77791-2
3	ddc90b51-27b6...	...	...	..				SURVEY.HAQ	4				Over the past W, are yo...								HAQ	...	2017-09-22		75821-9
4	ddc90b51-27b6...	...	...	..				SURVEY.PROMIS	4				I have been able to add ...								PRO...	...	2011-07-27		88854-5
5	ddc90b51-27b6...	...	...	..				HEM/BC	1				Erythrocytes			RBC (Amn fld) [#/Vol]		10*3/uL	10*3/uL				2014-02-03		55779-3
6	87350d0c-1871...	...	...	..				CHEM	1				Urea nitrogen			Urea nitrogen (CSF) [...		mg/dL	mg/dL				2009-06-16		14000-4
7	87350d0c-1871...	...	...	..				DRUG/TOX	1				Benzoylecgonine cutoff			Benzoylecgonine cuto...		ng/mL	ng/mL				2017-09-05		77791-2
8	87350d0c-1871...	...	...	..				SURVEY.HAQ	4				Over the past W, are yo...								HAQ	...	2017-09-22		75821-9
9	87350d0c-1871...	...	...	..				SURVEY.PROMIS	4				I have been able to add ...								PRO...	...	2011-07-27		88854-5
10	87350d0c-1871...	...	...	..				HEM/BC	1				Erythrocytes			RBC (Amn fld) [#/Vol]		10*3/uL	10*3/uL				2014-02-03		55779-3
11	87350d0c-1871...	...	...	..				H&P.HX	2				Braden scale score.total					{score}	score		Braden	...	1991-01-07		38227-5
12	87350d0c-1871...	...	...	..				MICRO	1				Yersinia pseudotubercul...			Y. pseudotuberculosis ...							1992-02-06		40936-7
13	87350d0c-1871...	...	...	..				MICRO	1				Human coronavirus OC4...		...	HCoV OC43 RNA NAA...							1994-12-05		82164-5
14	87350d0c-1871...	...	...	..				NIH.COGNITIVE	2				Can pronounce writhe								NIH_...	...	1989-03-01		84617-0

Display 5: SynthPHR™ - Lab Results

PROCEDURES

This functional block randomly assigns SNOMED-CT CORE-defined procedures to synthetic subjects.

The platform uses SNOMED-CT CORE’s Occurrence and Usage to generate random procedure records with the proper statistical frequency.

Display 6: SynthPHR™ - Procedures below shows the Procedures section of a SynthPHR™ SQLite file.

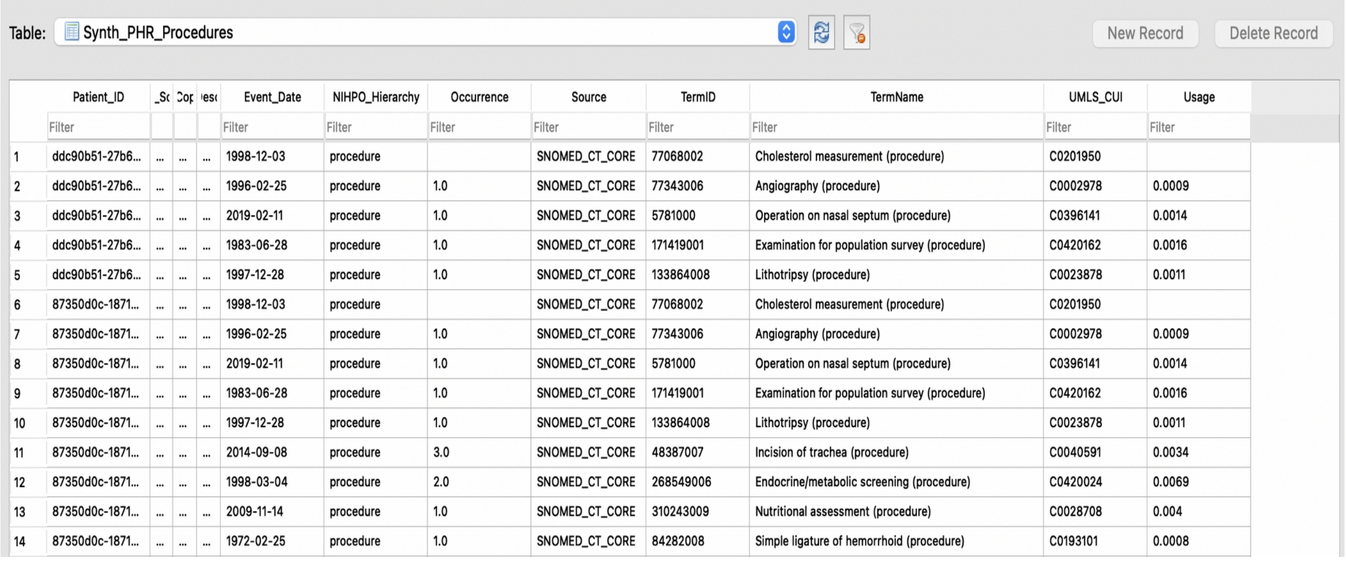


Table: Synth_PHR_Procedures

	Patient_ID	_Sc	Cop	)esc	Event_Date	NIHPO_Hierarchy	Occurrence	Source	TermID	TermName	UMLS_CUI	Usage
	Filter				Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	ddc90b51-27b6...	...	...	...	1998-12-03	procedure		SNOMED_CT_CORE	77068002	Cholesterol measurement (procedure)	C0201950	
2	ddc90b51-27b6...	...	...	...	1996-02-25	procedure	1.0	SNOMED_CT_CORE	77343006	Angiography (procedure)	C0002978	0.0009
3	ddc90b51-27b6...	...	...	...	2019-02-11	procedure	1.0	SNOMED_CT_CORE	5781000	Operation on nasal septum (procedure)	C0396141	0.0014
4	ddc90b51-27b6...	...	...	...	1983-06-28	procedure	1.0	SNOMED_CT_CORE	171419001	Examination for population survey (procedure)	C0420162	0.0016
5	ddc90b51-27b6...	...	...	...	1997-12-28	procedure	1.0	SNOMED_CT_CORE	133864008	Lithotripsy (procedure)	C0023878	0.0011
6	87350d0c-1871...	...	...	...	1998-12-03	procedure		SNOMED_CT_CORE	77068002	Cholesterol measurement (procedure)	C0201950	
7	87350d0c-1871...	...	...	...	1996-02-25	procedure	1.0	SNOMED_CT_CORE	77343006	Angiography (procedure)	C0002978	0.0009
8	87350d0c-1871...	...	...	...	2019-02-11	procedure	1.0	SNOMED_CT_CORE	5781000	Operation on nasal septum (procedure)	C0396141	0.0014
9	87350d0c-1871...	...	...	...	1983-06-28	procedure	1.0	SNOMED_CT_CORE	171419001	Examination for population survey (procedure)	C0420162	0.0016
10	87350d0c-1871...	...	...	...	1997-12-28	procedure	1.0	SNOMED_CT_CORE	133864008	Lithotripsy (procedure)	C0023878	0.0011
11	87350d0c-1871...	...	...	...	2014-09-08	procedure	3.0	SNOMED_CT_CORE	48387007	Incision of trachea (procedure)	C0040591	0.0034
12	87350d0c-1871...	...	...	...	1998-03-04	procedure	2.0	SNOMED_CT_CORE	268549006	Endocrine/metabolic screening (procedure)	C0420024	0.0069
13	87350d0c-1871...	...	...	...	2009-11-14	procedure	1.0	SNOMED_CT_CORE	310243009	Nutritional assessment (procedure)	C0028708	0.004
14	87350d0c-1871...	...	...	...	1972-02-25	procedure	1.0	SNOMED_CT_CORE	84282008	Simple ligature of hemorrhoid (procedure)	C0193101	0.0008

Display 6: SynthPHR™ - Procedures

PROVIDERS

This functional block uses Real-World Data from Medicare’s NPI Registry (06.6 Million registered Medicare providers) to randomly assign healthcare providers to each synthetic subject’s record.

Display 7: SynthPHR™ - Providers below shows the Providers section of a SynthPHR™ SQLite file.

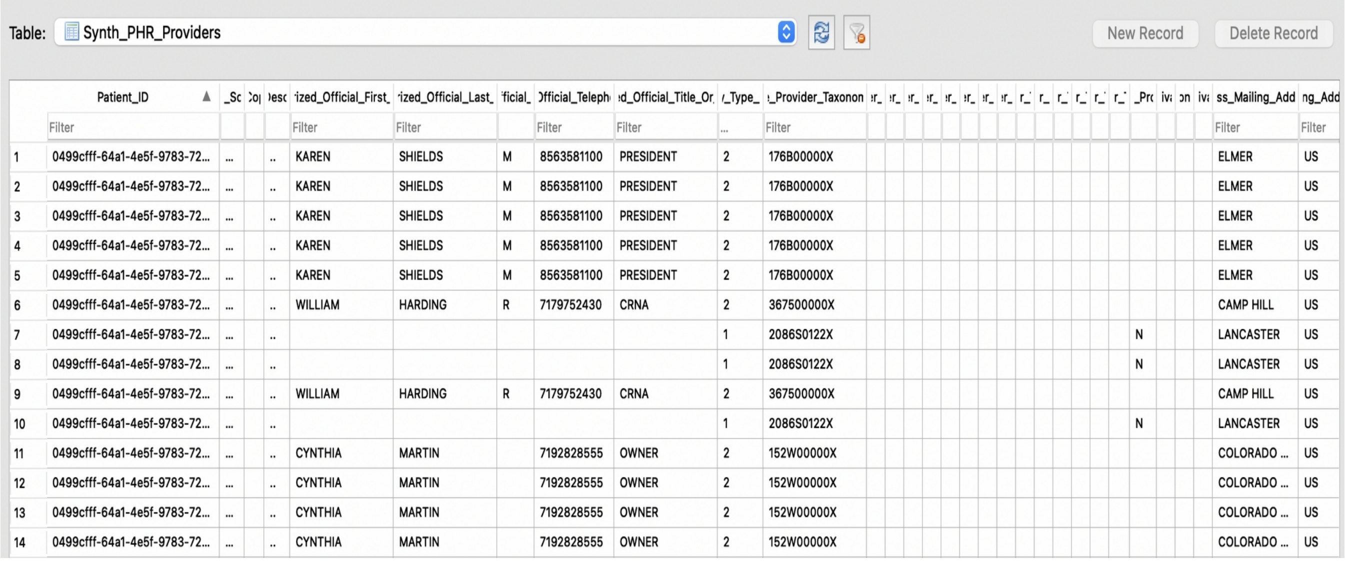


Table: Synth_PHR_Providers

	Patient_ID ▲	_Sc	Co|	)esc	ized_Official_First_	ized_Official_Last_	ficial_	Official_Teleph	d_Official_Title_Or	_Type_	_Provider_Taxonon	ır_	ır_	ır_	ır_	ır_	ır_	ır_	r_'	r_	r_'	r_'	r_'	r_'	_Prc	ivi	on	ivi	ss_Mailing_Add	ng_Add
	Filter				Filter	Filter		Filter	Filter	...	Filter																		Filter	Filter
1	0499cfff-64a1-4e5f-9783-72...	...		..	KAREN	SHIELDS	M	8563581100	PRESIDENT	2	176B00000X																		ELMER	US
2	0499cfff-64a1-4e5f-9783-72...	...		..	KAREN	SHIELDS	M	8563581100	PRESIDENT	2	176B00000X																		ELMER	US
3	0499cfff-64a1-4e5f-9783-72...	...		..	KAREN	SHIELDS	M	8563581100	PRESIDENT	2	176B00000X																		ELMER	US
4	0499cfff-64a1-4e5f-9783-72...	...		..	KAREN	SHIELDS	M	8563581100	PRESIDENT	2	176B00000X																		ELMER	US
5	0499cfff-64a1-4e5f-9783-72...	...		..	KAREN	SHIELDS	M	8563581100	PRESIDENT	2	176B00000X																		ELMER	US
6	0499cfff-64a1-4e5f-9783-72...	...		..	WILLIAM	HARDING	R	7179752430	CRNA	2	367500000X																		CAMP HILL	US
7	0499cfff-64a1-4e5f-9783-72...	...		..						1	2086S0122X														N				LANCASTER	US
8	0499cfff-64a1-4e5f-9783-72...	...		..						1	2086S0122X														N				LANCASTER	US
9	0499cfff-64a1-4e5f-9783-72...	...		..	WILLIAM	HARDING	R	7179752430	CRNA	2	367500000X																		CAMP HILL	US
10	0499cfff-64a1-4e5f-9783-72...	...		..						1	2086S0122X														N				LANCASTER	US
11	0499cfff-64a1-4e5f-9783-72...	...		..	CYNTHIA	MARTIN		7192828555	OWNER	2	152W00000X																		COLORADO ...	US
12	0499cfff-64a1-4e5f-9783-72...	...		..	CYNTHIA	MARTIN		7192828555	OWNER	2	152W00000X																		COLORADO ...	US
13	0499cfff-64a1-4e5f-9783-72...	...		..	CYNTHIA	MARTIN		7192828555	OWNER	2	152W00000X																		COLORADO ...	US
14	0499cfff-64a1-4e5f-9783-72...	...		..	CYNTHIA	MARTIN		7192828555	OWNER	2	152W00000X																		COLORADO ...	US

Display 7: SynthPHR™ - Providers

VITALS:

This functional block randomly assigns realistic vital sign readings to the synthetic person's record.

Display 8: SynthPHR™ - Vitals below shows the Vitals section of a SynthPHR™ SQLite file.

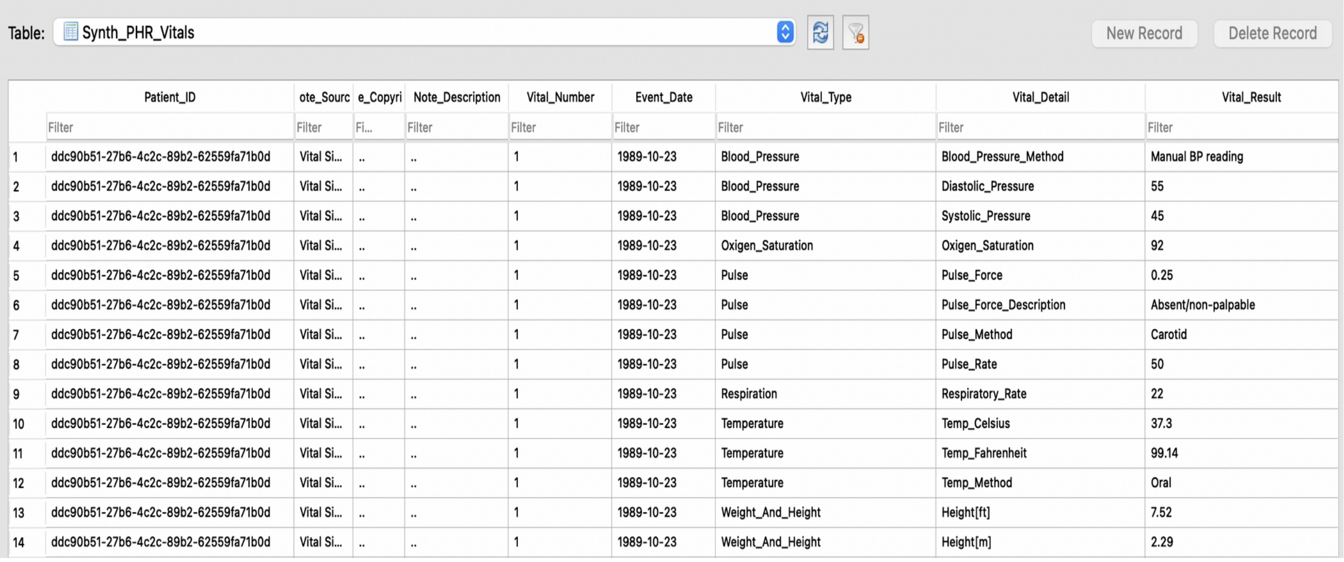


Table: Synth_PHR_Vitals

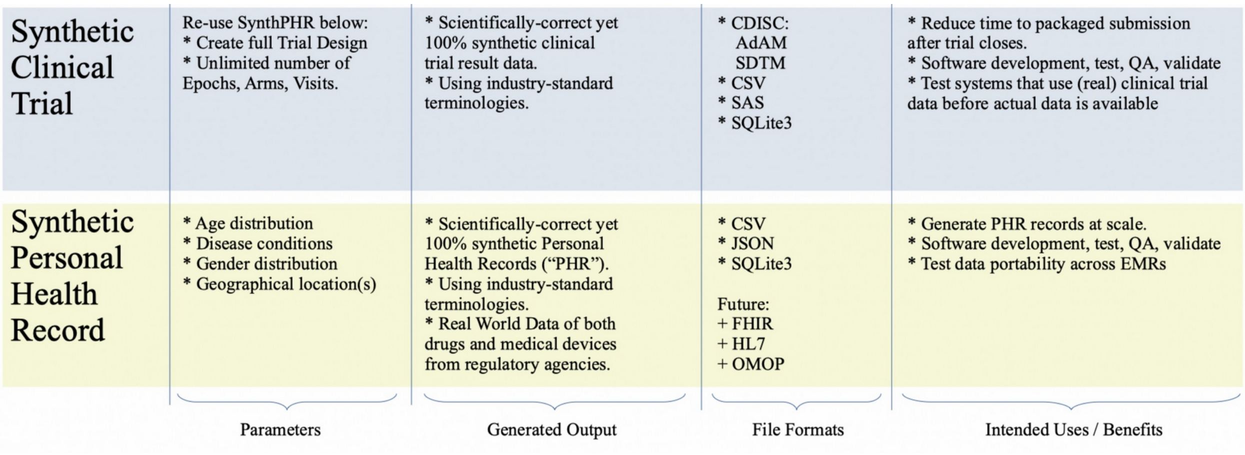
	Patient_ID	ote_Sourc	e_Copyri	Note_Description	Vital_Number	Event_Date	Vital_Type	Vital_Detail	Vital_Result
	Filter	Filter	Fi...	Filter	Filter	Filter	Filter	Filter	Filter
1	ddc90b51-27b6-4c2c-89b2-62559fa71b0d	Vital Si...	..	..	1	1989-10-23	Blood_Pressure	Blood_Pressure_Method	Manual BP reading
2	ddc90b51-27b6-4c2c-89b2-62559fa71b0d	Vital Si...	..	..	1	1989-10-23	Blood_Pressure	Diastolic_Pressure	55
3	ddc90b51-27b6-4c2c-89b2-62559fa71b0d	Vital Si...	..	..	1	1989-10-23	Blood_Pressure	Systolic_Pressure	45
4	ddc90b51-27b6-4c2c-89b2-62559fa71b0d	Vital Si...	..	..	1	1989-10-23	Oxigen_Saturation	Oxigen_Saturation	92
5	ddc90b51-27b6-4c2c-89b2-62559fa71b0d	Vital Si...	..	..	1	1989-10-23	Pulse	Pulse_Force	0.25
6	ddc90b51-27b6-4c2c-89b2-62559fa71b0d	Vital Si...	..	..	1	1989-10-23	Pulse	Pulse_Force_Description	Absent/non-palpable
7	ddc90b51-27b6-4c2c-89b2-62559fa71b0d	Vital Si...	..	..	1	1989-10-23	Pulse	Pulse_Method	Carotid
8	ddc90b51-27b6-4c2c-89b2-62559fa71b0d	Vital Si...	..	..	1	1989-10-23	Pulse	Pulse_Rate	50
9	ddc90b51-27b6-4c2c-89b2-62559fa71b0d	Vital Si...	..	..	1	1989-10-23	Respiration	Respiratory_Rate	22
10	ddc90b51-27b6-4c2c-89b2-62559fa71b0d	Vital Si...	..	..	1	1989-10-23	Temperature	Temp_Celsius	37.3
11	ddc90b51-27b6-4c2c-89b2-62559fa71b0d	Vital Si...	..	..	1	1989-10-23	Temperature	Temp_Fahrenheit	99.14
12	ddc90b51-27b6-4c2c-89b2-62559fa71b0d	Vital Si...	..	..	1	1989-10-23	Temperature	Temp_Method	Oral
13	ddc90b51-27b6-4c2c-89b2-62559fa71b0d	Vital Si...	..	..	1	1989-10-23	Weight_And_Height	Height[ft]	7.52
14	ddc90b51-27b6-4c2c-89b2-62559fa71b0d	Vital Si...	..	..	1	1989-10-23	Weight_And_Height	Height[m]	2.29

Display 8: SynthPHR™ - Vitals

Please note that in all functional blocks listed above, every date in every event (procedures, taking drugs, inserted devices, etc.) falls within the assigned lifetime of the synthetic subject.

BUILDING A SYNTHETIC CLINICAL TRIAL (SYNTHTRIAL™)

The third stage in the platform is to create a synthetic clinical trial using the synthetic cohort defined above. Figure 2: Synthetic Clinical Trial (SynthTrial™) overview below summarizes the goals and objectives of a synthetic trial.

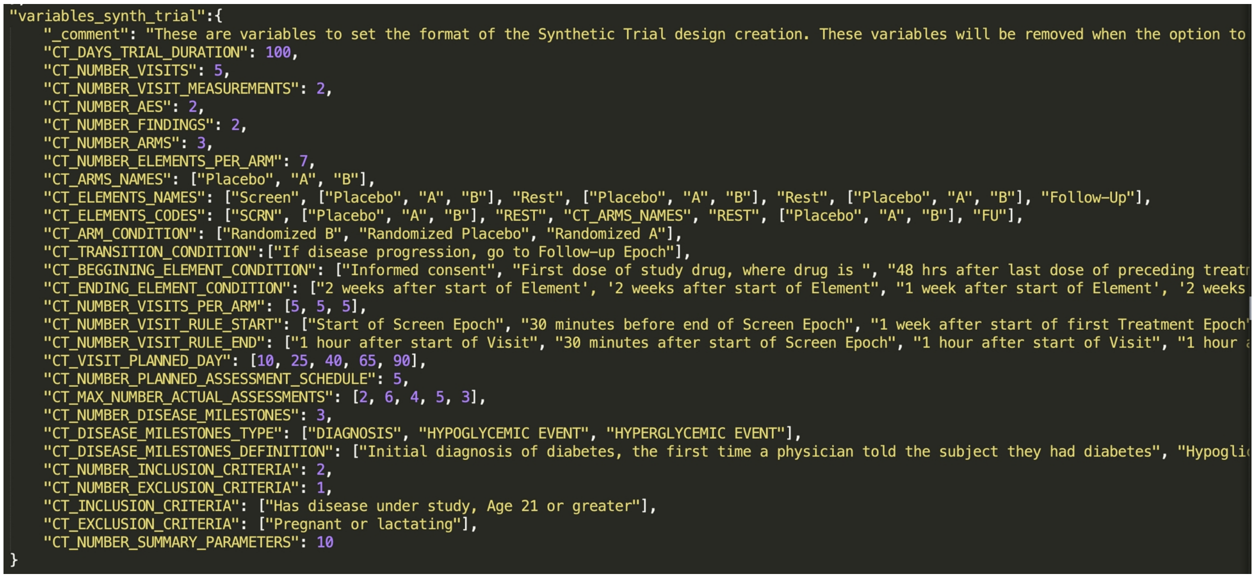


	Parameters	Generated Output	File Formats	Intended Uses / Benefits
Synthetic Clinical Trial	Re-use SynthPHR below: * Create full Trial Design * Unlimited number of Epochs, Arms, Visits.	* Scientifically-correct yet 100% synthetic clinical trial result data. * Using industry-standard terminologies.	* CDISC: AdAM SDTM * CSV * SAS * SQLite3	* Reduce time to packaged submission after trial closes. * Software development, test, QA, validate * Test systems that use (real) clinical trial data before actual data is available
Synthetic Personal Health Record	* Age distribution * Disease conditions * Gender distribution * Geographical location(s)	* Scientifically-correct yet 100% synthetic Personal Health Records ("PHR"). * Using industry-standard terminologies. * Real World Data of both drugs and medical devices from regulatory agencies.	* CSV * JSON * SQLite3 Future: + FHIR + HL7 + OMOP	* Generate PHR records at scale. * Software development, test, QA, validate * Test data portability across EMRs

Figure 2: Synthetic Clinical Trial (SynthTrial™) overview

We'll now describe the process the platform follows to generate a full synthetic trial. First, the software takes the synthetic cohort defined above. Then the user defines all the parameters specified in a Trial Design form.

Figure 3: Trial Design form below is a sample representation of how a user defines the parameters of a synthetic clinical trial. There is also a user-friendly GUI available to enter the Trial Design parameters.



```
"variables_synth_trial":{
    "_comment": "These are variables to set the format of the Synthetic Trial design creation. These variables will be removed when the option to
    "CT_DAYS_TRIAL_DURATION": 100,
    "CT_NUMBER_VISITS": 5,
    "CT_NUMBER_VISIT_MEASUREMENTS": 2,
    "CT_NUMBER_AES": 2,
    "CT_NUMBER_FINDINGS": 2,
    "CT_NUMBER_ARMS": 3,
    "CT_NUMBER_ELEMENTS_PER_ARM": 7,
    "CT_ARMS_NAMES": ["Placebo", "A", "B"],
    "CT_ELEMENTS_NAMES": ["Screen", ["Placebo", "A", "B"], "Rest", ["Placebo", "A", "B"], "Rest", ["Placebo", "A", "B"], "Follow-Up"],
    "CT_ELEMENTS_CODES": ["SCRN", ["Placebo", "A", "B"], "REST", "CT_ARMS_NAMES", "REST", ["Placebo", "A", "B"], "FU"],
    "CT_ARM_CONDITION": ["Randomized B", "Randomized Placebo", "Randomized A"],
    "CT_TRANSITION_CONDITION":["If disease progression, go to Follow-up Epoch"],
    "CT_BEGGINING_ELEMENT_CONDITION": ["Informed consent", "First dose of study drug, where drug is ", "48 hrs after last dose of preceding treatm
    "CT_ENDING_ELEMENT_CONDITION": ["2 weeks after start of Element', '2 weeks after start of Element", "1 week after start of Element', '2 weeks
    "CT_NUMBER_VISITS_PER_ARM": [5, 5, 5],
    "CT_NUMBER_VISIT_RULE_START": ["Start of Screen Epoch", "30 minutes before end of Screen Epoch", "1 week after start of first Treatment Epoch"
    "CT_NUMBER_VISIT_RULE_END": ["1 hour after start of Visit", "30 minutes after start of Screen Epoch", "1 hour after start of Visit", "1 hour a
    "CT_VISIT_PLANNED_DAY": [10, 25, 40, 65, 90],
    "CT_NUMBER_PLANNED_ASSESSMENT_SCHEDULE": 5,
    "CT_MAX_NUMBER_ACTUAL_ASSESSMENTS": [2, 6, 4, 5, 3],
    "CT_NUMBER_DISEASE_MILESTONES": 3,
    "CT_DISEASE_MILESTONES_TYPE": ["DIAGNOSIS", "HYPOGLYCEMIC EVENT", "HYPERGLYCEMIC EVENT"],
    "CT_DISEASE_MILESTONES_DEFINITION": ["Initial diagnosis of diabetes, the first time a physician told the subject they had diabetes", "Hypoglic
    "CT_NUMBER_INCLUSION_CRITERIA": 2,
    "CT_NUMBER_EXCLUSION_CRITERIA": 1,
    "CT_INCLUSION_CRITERIA": ["Has disease under study, Age 21 or greater"],
    "CT_EXCLUSION_CRITERIA": ["Pregnant or lactating"],
    "CT_NUMBER_SUMMARY_PARAMETERS": 10
}
```

Figure 3: Trial Design form

Once the user defines the desired parameters, the platform then proceeds to generate the trial data following these steps.

01. TABLE-DRIVEN RECORD CREATION

NIHPO has codified the CDISC SDTM structure into a series of tables in a SQLite database file. Think of this as the “Computable CDISC Standards.”

The platform uses these tables to programmatically create a record for each SDTM Domain, where the Domain record structure (name, description, and type of field) is defined in these tables.

Display 9: CDISC SDTM structure definition below shows how the platform represents the CDISC SDTM Domains in a customized table.

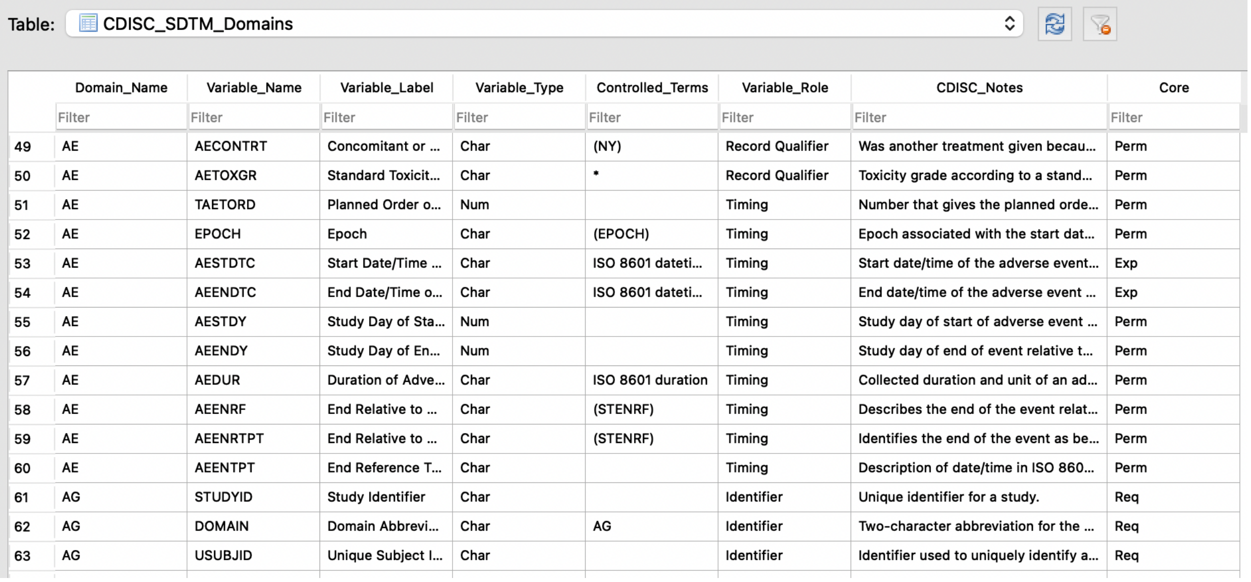


Table: CDISC_SDTM_Domains

	Domain_Name	Variable_Name	Variable_Label	Variable_Type	Controlled_Terms	Variable_Role	CDISC_Notes	Core
	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
49	AE	AECONTRT	Concomitant or ...	Char	(NY)	Record Qualifier	Was another treatment given becau...	Perm
50	AE	AETOXGR	Standard Toxicit...	Char	*	Record Qualifier	Toxicity grade according to a stand...	Perm
51	AE	TAETORD	Planned Order o...	Num		Timing	Number that gives the planned orde...	Perm
52	AE	EPOCH	Epoch	Char	(EPOCH)	Timing	Epoch associated with the start dat...	Perm
53	AE	AESTDTC	Start Date/Time ...	Char	ISO 8601 dateti...	Timing	Start date/time of the adverse event...	Exp
54	AE	AEENDTC	End Date/Time o...	Char	ISO 8601 dateti...	Timing	End date/time of the adverse event ...	Exp
55	AE	AESTDY	Study Day of Sta...	Num		Timing	Study day of start of adverse event ...	Perm
56	AE	AEENDY	Study Day of En...	Num		Timing	Study day of end of event relative t...	Perm
57	AE	AEDUR	Duration of Adve...	Char	ISO 8601 duration	Timing	Collected duration and unit of an ad...	Perm
58	AE	AEENRF	End Relative to ...	Char	(STENRF)	Timing	Describes the end of the event relat...	Perm
59	AE	AEENRTPT	End Relative to ...	Char	(STENRF)	Timing	Identifies the end of the event as be...	Perm
60	AE	AEENTPT	End Reference T...	Char		Timing	Description of date/time in ISO 860...	Perm
61	AG	STUDYID	Study Identifier	Char		Identifier	Unique identifier for a study.	Req
62	AG	DOMAIN	Domain Abbrevi...	Char	AG	Identifier	Two-character abbreviation for the ...	Req
63	AG	USUBJID	Unique Subject I...	Char		Identifier	Identifier used to uniquely identify a...	Req

Display 9: CDISC SDTM structure definition

This approach of representing the Domains in a database table allows developers to quickly and easily extend the platform’s functionality to other CDISC standards (AdAM is next on our list). And developer can also quickly adapt to changes in the standards: developer will simply update this table to reflect any changes to Domain-specific fields.

02. PROCESSING RULES DEFINED IN TABLES

Before the platform generates a record for a Domain the software needs to know the “creation rules” for each Domain.

For example: for the “AE” Domain, the platform needs to create one record per Subject, per Adverse Event. These “rules” are codified in a database table in the SQLite file as well.

Display 10: CDISC SDTM Domain Rules below displays part of the rules defining what records to create for each Domain.

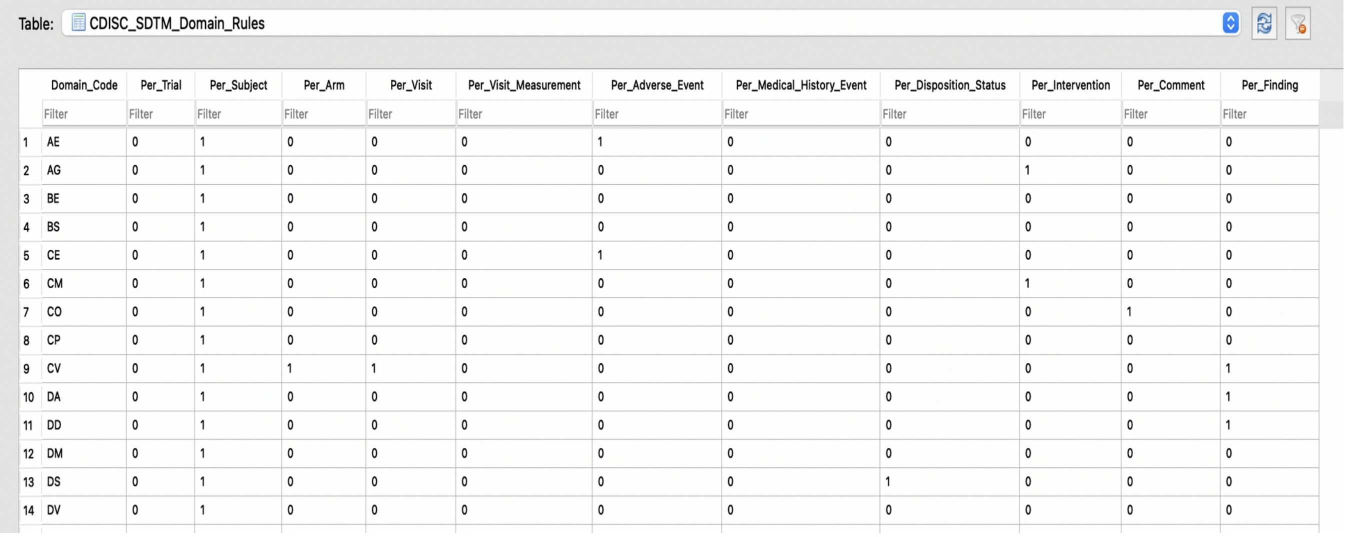


Table: CDISC_SDTM_Domain_Rules

	Domain_Code	Per_Trial	Per_Subject	Per_Arm	Per_Visit	Per_Visit_Measurement	Per_Adverse_Event	Per_Medical_History_Event	Per_Disposition_Status	Per_Intervention	Per_Comment	Per_Finding
	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	AE	0	1	0	0	0	1	0	0	0	0	0
2	AG	0	1	0	0	0	0	0	0	1	0	0
3	BE	0	1	0	0	0	0	0	0	0	0	0
4	BS	0	1	0	0	0	0	0	0	0	0	0
5	CE	0	1	0	0	0	1	0	0	0	0	0
6	CM	0	1	0	0	0	0	0	0	1	0	0
7	CO	0	1	0	0	0	0	0	0	0	1	0
8	CP	0	1	0	0	0	0	0	0	0	0	0
9	CV	0	1	1	1	0	0	0	0	0	0	1
10	DA	0	1	0	0	0	0	0	0	0	0	1
11	DD	0	1	0	0	0	0	0	0	0	0	1
12	DM	0	1	0	0	0	0	0	0	0	0	0
13	DS	0	1	0	0	0	0	0	1	0	0	0
14	DV	0	1	0	0	0	0	0	0	0	0	0

Display 10: CDISC SDTM Domain Rules

This table can be easily updated as the standards evolve.

03. RE-USE FUNCTIONAL BLOCKS

The platform now knows what fields to create for each Domain record. And the rules for when creating each record. Next, the platform has to “fill-in” each Domain-specific record. For example, the “LB” Domain has the fields "LBLOINC" ("LOINC Code") and "LBSPEC" ("Specimen Type").

As seen in Table 2. Functional Blocks used to generate Personal Health Records above, the platform generates data of different types on demand. To populate the LB Domain, the platform calls the Lab Results functional block and receives appropriate values to insert into the LBLOINC and LBSPEC fields.

And in the BS Domain, the fields BSTESTCD, BSTEST, BSORRESCU, and BSSTRESU are populated with values from the CDISC Controlled Terminology. See Display 11: CDISC SDTM BS Domain below.

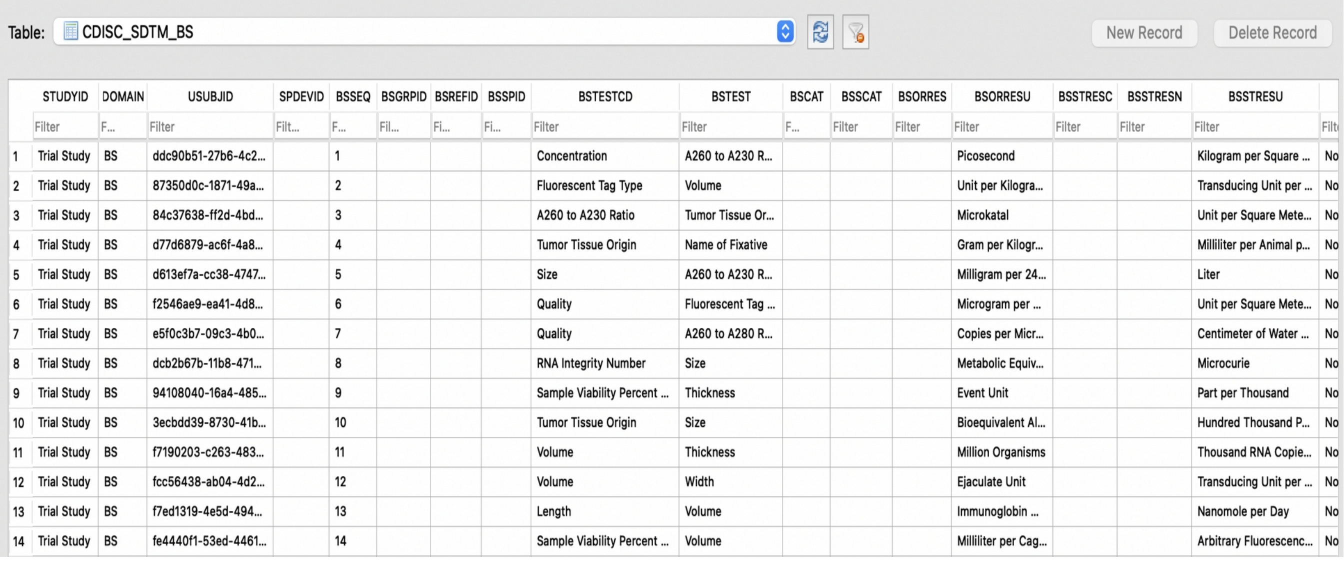


Table: CDISC_SDTM_BS

	STUDYID	DOMAIN	USUBJID	SPDEVID	BSSEQ	BSGRPID	BSREFID	BSSPID	BSTESTCD	BSTEST	BSCAT	BSSCAT	BSORRES	BSORRESU	BSSTRESC	BSSTRESN	BSSTRESU	
	Filter	F...	Filter	Filt...	F...	Fil...	Fi...	Fi...	Filter	Filter	F...	Filter	Filter	Filter	Filter	Filter	Filter	Filt
1	Trial Study	BS	ddc90b51-27b6-4c2...		1				Concentration	A260 to A230 R...				Picosecond			Kilogram per Square ...	No
2	Trial Study	BS	87350d0c-1871-49a...		2				Fluorescent Tag Type	Volume				Unit per Kilogra...			Transducing Unit per ...	No
3	Trial Study	BS	84c37638-ff2d-4bd...		3				A260 to A230 Ratio	Tumor Tissue Or...				Microkatal			Unit per Square Mete...	No
4	Trial Study	BS	d77d6879-ac6f-4a8...		4				Tumor Tissue Origin	Name of Fixative				Gram per Kilogr...			Milliliter per Animal p...	No
5	Trial Study	BS	d613ef7a-cc38-4747...		5				Size	A260 to A230 R...				Milligram per 24...			Liter	No
6	Trial Study	BS	f2546ae9-ea41-4d8...		6				Quality	Fluorescent Tag ...				Microgram per ...			Unit per Square Mete...	No
7	Trial Study	BS	e5f0c3b7-09c3-4b0...		7				Quality	A260 to A280 R...				Copies per Micr...			Centimeter of Water ...	No
8	Trial Study	BS	dcb2b67b-11b8-471...		8				RNA Integrity Number	Size				Metabolic Equiv...			Microcurie	No
9	Trial Study	BS	94108040-16a4-485...		9				Sample Viability Percent ...	Thickness				Event Unit			Part per Thousand	No
10	Trial Study	BS	3ecbdd39-8730-41b...		10				Tumor Tissue Origin	Size				Bioequivalent Al...			Hundred Thousand P...	No
11	Trial Study	BS	f7190203-c263-483...		11				Volume	Thickness				Million Organisms			Thousand RNA Copie...	No
12	Trial Study	BS	fcc56438-ab04-4d2...		12				Volume	Width				Ejaculate Unit			Transducing Unit per ...	No
13	Trial Study	BS	f7ed1319-4e5d-494...		13				Length	Volume				Immunoglobin ...			Nanomole per Day	No
14	Trial Study	BS	fe4440f1-53ed-4461...		14				Sample Viability Percent ...	Volume				Milliliter per Cag...			Arbitrary Fluorescenc...	No

Display 11: CDISC SDTM BS Domain

The USUBJID field in each SDTM record corresponds to the Subject ID in the synthetic PHR. User can trace a single synthetic subject throughout the clinical trial and all the way back to the subject's personal health history.

Please notice that this modular approach to generating and re-using synthetic health data is highly scalable: developer can add new functional blocks (e.g, generate synthetic medical images) and the platform will call those new functional blocks as needed (e.g. populate an EDC with synthetic images).

04. PANDAS-BASED INTERNAL STRUCTURE REPRESENTATION

The entire platform runs as a series of Pandas DataFrames for high performance and ease of use.

The use of Pandas also gives the platform enormous flexibility and scalability. The internal representation of the Domains is easily changeable on the fly, and the Pandas DataFrames can be exported to a large variety of output formats.

05. FLEXIBLE OUTPUT

The platform currently exports all CDISC SDTM domains in CSV, SAS, and SQLite formats. Developers can add the ability to generate new output format types (FHIR, HL7, OMOP) based on customer needs.

NIHPO already demonstrated how the platform can easily load the synthetic data (both PHR as well as trial data) directly into an Electronic Data Capture ("EDC") system like Medidata Rave.

06. CONFORMANCE LEVELS

The platform's output files are designed to meet several conformance levels. In order of increasing complexity:

- Structural conformance: All the generated output files have a structure that matches the expected structure of the SAS XPORT format.
- Standard (CDISC) conformance: The format of each generated record in each SDTM Domain matches the CDISC definitions. The output SAS files generated by the platform are validated with the Pinnacle 21 Community edition.
- Scientific conformance: The platform uses both Controlled Terminologies as well as Real-World Data to ensure the content of each record, while randomly-generated, is scientifically accurate.
- Semantic conformance: The output files are not intended to replace "real" healthcare data. Rather, the goal of this synthetic data is to act as a temporary placeholder for real data. Therefore, the generated synthetic data files are not intended to be semantically correct.

Synthea's output, on the other hand, has a high degree of both scientific and semantic compliance due to Synthea's use of their fantastic Generic Module Framework (GMF) that "enables the modeling of various diseases and conditions that contribute to the medical history of synthetic patients."

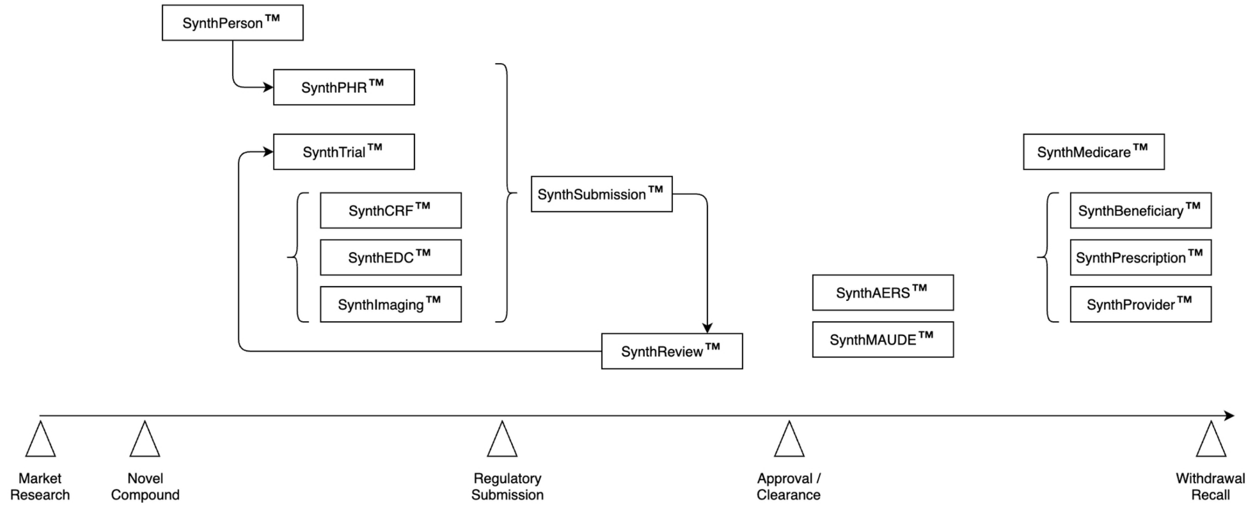
FUTURE WORK AND EXTENSIONS

NIHPO's long-term goal is to build a "Clinical Trial Composer." This will be a desktop application where users will be able to "compose" a Clinical Trial just like a musician composes a song, or a movie director edits raw video footage, on their desktops.

Musicians use sophisticated GUIs to lay tracks, edit sequences and audio samples, and then to generate the finished product. We wonder why not using the same mental model for "composing" (adding, editing, aligning, testing, outputting) all the elements of a (synthetic) clinical trial?

The Clinical Trial Composer will enable any staff role (from a market-focused analyst to a data-focused statistician) to quickly and easily define, create, and use synthetic health data. Either instead of or as a placeholder until real data is available. From the user's desktop, without IT involvement.

Drawing 1: Synthetic Health Data roadmap below shows NIHPO's roadmap for synthetic health data.



Drawing 1: Synthetic Health Data roadmap

Synthetic Case Report Forms (SynthCRF™)

The platform already generates valid records for Adverse Events, for example. It is then easy to extend the platform to fill out the corresponding forms for Adverse Events. The same method applies to all forms.

The platform will populate each trial-related form based on a user-defined mapping of what fields go where in each form.

Synthetic Electronic Data Capture (SynthEDC™)

A logical output for the platform is to populate an EDC (Electronic Data Capture) system with all the synthetic data generated for both subjects as well as the trial events, visits, measurements.

At this time the platform can already programmatically populate Medidata Rave through Rave's API.

Synthetic Medical Imaging (SynthImaging™)

The platform will be extended to generate on-demand synthetic medical images at scale. The platform will use publicly-available medical imaging datasets as a reference for all imaging modalities and disease conditions. The platform will then randomly insert markers into the reference images to create synthetic images. For example: the platform will generate breast cancer images with randomly-placed lesions.

These medical images will be customizable to fit a user's specific needs. The synthetic image files will be unencumbered, scientifically valid, and available in multiple formats (DICOM, JPG, PNG).

CONCLUSION

This paper described a database-driven software platform that generates realistic, yet synthetic health data. The use of both Real-World Data and controlled terminologies ensures that the generated data is scientifically valid. The software is structured as functional blocks that are easily re-used and extensible.

We presented additional functionality and a road map of future capabilities under development.

The platform's source code is available in GitHub at https://github.com/nihpo/SynthHealthData under an Open Source license (AGPL 3.0).

TOOLS

These are the Python libraries used in this work:

Pandas [https://pandas.pydata.org/]

Xport [https://github.com/selik/xport]

REFERENCES

Center for Medicare and Medicaid Services ("CMS") National Provider Identifier. Accessed 13 April 2021. https://npiregistry.cms.hhs.gov/

European Medicines Agency ("EMA") Accessed 13 April 2021. https://www.ema.europa.eu/en/medicines/download-medicine-data

Food and Drug Administration's ("FDA") FDA's Drugs@FDA Accessed 13 April 2021. https://www.accessdata.fda.gov/scripts/cder/daf/

GIS. A geographic information system (GIS) is a conceptualized framework that provides the ability to capture and analyze spatial and geographic data. Accessed 13 April 2021. https://en.wikipedia.org/wiki/Geographic_information_system

PHUSE Test Data Factory working group. Accessed 13 April 2021. https://github.com/phuse-org/TestDataFactory

Pinnacle 21 Community. Accessed April 13, 2021. https://www.pinnacle21.com/downloads

Replica Analytics. Replica Analytics provides solutions to generate synthetic data based on real datasets, and/or perform privacy assurance on the generated data, regardless of where and by whom the data was generated. Accessed April 13, 2021. https://replica-analytics.com/home

SQLite. SQLite is a C-language library that implements a small, fast, self-contained, high-reliability, full-featured, SQL database engine. Accessed April 13, 2021. https://sqlite.org/index.html

Synthea. Synthea™ is an open-source, synthetic patient generator that models the medical history of synthetic patients. Accessed April 13, 2021. https://synthetichealth.github.io/synthea/

US Locations. U.S. Geological Survey's National Geospatial Program; U.S. Board on Geographic Names. Accessed April 13, 2021. https://www.usgs.gov/core-science-systems/ngp/board-on-geographic-names/download-gnis-data

World Locations. National Geospatial Intelligence Agency, Complete Files of Geographic Names for Geopolitical Areas. Accessed April 13, 2021. https://geonames.nga.mil/gns/html/namefiles.html

ACKNOWLEDGMENTS

Author acknowledges the constant guidance and support from the PhUSE Test Data Factory (TDF) project. In particular: Dante Di Tommaso and Peter Schaefer. As well as: Nancy Brucken, Jessica Dai, Alexey Kuznetsov, and Cynthia Stroupe.

CONTACT INFORMATION

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

José C. Lacal, CTO
NIHPO, Inc.
+1 (561) 777-2577 Jose.Lacal@NIHPO.com
http://NIHPO.com and https://github.com/nihpo/SynthHealthData