Tanzania

Synthetic Data for ALPHA Residencies and HIV Testing

Study Documentation

December 14, 2022

Metadata Production

Metadata Producer(s)	Tathagata Bhattacharjee (TB) , LSHTM , Synthetic Data Producer & DDI Author
Identification	DDI.INSPIRE.SYNTHETIC.ALPHA.V1.0

Table of Contents

Synthetic Data for ALPHA Residencies and HIV Testing

Overview

Identification	INSPIRE.SYNTHETIC.ALPHA.V1.0
Version	INSPIRE.SYNTHETIC.ALPHA.V1.0 - First public release of synthetic data simulating the ALPHA residencies and HIV Testing data specifications

Abstract

This synthetic datasets has been created based on the ALPHA data specifications.

The ALPHA Network is a collaboration between 10 longitudinal studies in sub-Saharan Africa. These studies collect data on HIV infection alongside demographic, behavioural, socio-economic and clinical data from residents of the study areas. The Network harmonises these data and conducts comparable and pooled analyses on HIV-related research questions.

The synthetic dataset consists of the following data specificatons:

Residency (formerly called 6.1) contains date of birth, the periods of time spend resident in the study and how these ended. This information is used for survival analysis. Residency data spec definition. [https://alpha.lshtm.ac.uk/wp-content/uploads//2022/06/Spec1_ALPHA_data_spec_residency_2019_04.pdf]

HIV tests (formerly called 6.2b) contains dates and results of HIV tests done for research purposes and, for some studies, information on tests done in other settings and self-reported HIV status. HIV tests data spec definition. [https://alpha.lshtm.ac.uk/wp-content/uploads//2022/06/Spec2_ALPHA_data_spec_HIV_tests_2019_04.pdf]

This synthetic data has been derived from the ALPHA data of The Kisesa Health and Demographic Surveillance System (Kisesa HDSS), which is part of Kisesa OpenCohort HIV Study located in a rural area of North-Western Tanzania. The process of synthetic data generation is explaind in "Datasets" section of this document.

A copy of ALPHA datasets from Kisesa HDSS was used as input for scanning using OHDSI's opensource WhiteRabbit tool. After scanning, a fake dataset was generated using a feature of WhiteRabbit tool. Later, data was adjusted for date dependiencies using Pentaho Data Integration tool.

Unit of Analysis	Individual

Scope & Coverage

Topics	Demography, HIV Testing
Countries	Tanzania

Geographic Coverage

Synthetic data generated from Kisesa Health and Demographic Surveillance System (Kisesa HDSS)

Universe

Syhthetic datasets generated by randomly picking up data from source and anonymized so as to ensure that it does not reflect any real or near-to-real characteristics of the original datasets.

Producers & Sponsors

Other Producer(s)	Tathagata Bhattacharjee (TB) , LSHTM , Synthetic data producer & DDI author
Funding Agency/ies	INSPIRE Network

Sampling

Sampling Procedure Synthetic data

Accessibility	
Distributor(s)	INSPIRE Network

Files Description

Dataset contains 2 file(s)

synthetic_residencies_v1_0	
# Cases	30000
# Variable(s)	17
File Structure	Type: relational Key(s): idno (idno)

File Content

ALPHA Network data specification: residency episodes

Residency episodes
• This is the starting point for all ALPHA analyses and is therefore essential for all study sites.
• It is used to compute person-year denominators for age-specific rates.
• We expect to see one record per episode of household residence for each individual in the data set, i.e. on average more than one record per person
• Only those individuals who have been resident continuously in the same household between first and last date of observation will have only one residence episode record.
• Individuals who have moved household within the DSS area, or have left and returned to the DSS area since the time they were first seen, will have two or more records, depending on the number of periods of absence from the study area and the number of times they have moved household. In this case, entry and exit dates and types in the classification below refer to the start and end of an episode of residence rather than to the first and last encounter with the individual in the study throughout the whole of his/her life.
• For records relating to consecutive residence episodes, where an individual moves within the study area, the (household) exit date of the earlier episode should be equal to the (household) entry date of the later episode.
• For records relating to individuals who moved out of the study area and then moved back in, the entry date of the later episode must be strictly greater than the exit date of the earlier episode.

This synthetic data has been derived from the ALPHA data of The Kisesa Health and Demographic Surveillance System (Kisesa HDSS), which is part of Kisesa OpenCohort HIV Study located in a rural area of North-Western Tanzania. The process of synthetic data generation is explaind in "Datasets" section of this document.

A copy of ALPHA datasets from Kisesa HDSS was used as input for scanning using OHDSI's opensource WhiteRabbit tool. After scanning, a fake dataset was generated using a feature of WhiteRabbit tool. Later, data was adjusted for date dependiencies using Pentaho Data Integration tool.

Producer

Tathagata Bhattacharjee, London School of Hygiene & Tropical Medicine (LSHTM) for INSPIRE Network (https://inspiredata.network/)

Version

ALPHA Residencies: ALPHA_SYNTHETIC_V1_0

synthetic_hiv_testing_v1_0	
# Cases	15000
# Variable(s)	10
File Structure	Type: relational Key(s): idno (idno)

File Content

ALPHA Network data specification: HIV tests

HIV test data
• This dataset is for the results of HIV tests from research activities or clinic data
• If your site uses self reports of HIV status in your testing procedure you should include them here
• For example, if the respondent discloses they are positive and you then do not test them, you should include the positive result in this spec and you record this under
source_of_test_information=3 “self-reported by respondent”.
• Sites that use the result of a recent test, carried out for research purposes shortly before the population based research study data collection(E.g. Kisumu and Rakai) should include both results
• once in “2 part of special research study” with test_assumption as “0 new test” • again in “1 part of a population based study” with test_assumption as “test from previous study used.
• Sites whose protocol says “do not test if person has previously tested positive” should record as though there was a positive test done on the date of study and put the test_assumption variable as “ 2 Previous positive HIV test used”.

This synthetic data has been derived from the ALPHA data of The Kisesa Health and Demographic Surveillance System (Kisesa HDSS), which is part of Kisesa OpenCohort HIV Study located in a rural area of North-Western Tanzania. The process of synthetic data generation is explaind in "Datasets" section of this document.

A copy of ALPHA datasets from Kisesa HDSS was used as input for scanning using OHDSI's opensource WhiteRabbit tool. After scanning, a fake dataset was generated using a feature of WhiteRabbit tool. Later, data was adjusted for date dependiencies using Pentaho Data Integration tool.

Producer
Tathagata Bhattacharjee, London School of Hygiene & Tropical Medicine (LSHTM) for INSPIRE Network (https://inspiredata.network/)

Version
ALPHA HIV Testing: ALPHA_SYNTHETIC_V1_0

Variables List

Dataset contains 27 variable(s)

File synthetic_residencies_v1_0

#	Name	Label	Type	Format	Valid	Invalid	Question
1	study_name	study_name	discrete	character-16	30000	0	-
2	idno	idno	continuous	numeric.0	30000	0	-
3	hhold_id	hhold_id	continuous	numeric.0	30000	0	-
4	hhold_id ..	hhold_id_extra	discrete	numeric.0	0	30000	-
5	sex	sex	discrete	numeric.0	30000	0	-
6	dob	dob	discrete	character	29981	-	-
7	residence	residence	discrete	numeric.0	30000	0	-
8	entry_type	entry_type	discrete	numeric.0	30000	0	-
9	entry_date	entry_date	discrete	character	29996	-	-
10	entry_ty ..	entry_type_of_date	discrete	numeric.0	29996	4	-
11	entry_ob ..	entry_obs_date	discrete	character	29996	-	-
12	entry_ob ..	entry_obs_round	continuous	numeric.0	30000	0	-
13	exit_type	exit_type	discrete	numeric.0	30000	0	-
14	exit_date	exit_date	discrete	character	29984	-	-
15	exit_typ ..	exit_type_of_date	discrete	numeric.0	29984	16	-
16	exit_obs ..	exit_obs_date	discrete	character	30000	-	-
17	exit_obs ..	exit_obs_round	continuous	numeric.0	30000	0	-

File synthetic_hiv_testing_v1_0

#	Name	Label	Type	Format	Valid	Invalid	Question
1	idno	idno	continuous	numeric.0	15000	0	-
2	study_name	study_name	discrete	character-16	15000	0	-
3	test_rep ..	test_report_date	discrete	character	15000	-	-
4	hiv_test ..	hiv_test_date	discrete	character	15000	-	-
5	hiv_test ..	hiv_test_result	discrete	numeric.0	11234	3766	-
6	informed ..	informed_of_result	discrete	numeric.0	15000	0	-
7	source_o ..	source_of_test_information	discrete	numeric.0	15000	0	-
8	test_ass ..	test_assumption	discrete	numeric.0	15000	0	-
9	original ..	original_hiv_test_result	discrete	numeric.0	0	15000	-
10	survey_r ..	survey_round_name	discrete	character-6	15000	0	-

Variables Description

Dataset contains 27 variable(s)

File : synthetic_residencies_v1_0

study_name: study_name

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=30000 /-] [Invalid=0 /-]
Definition	Name of the study field site. Character - consistent across data sets

Value	Label	Cases	Percentage
kisesa-synthetic		30000	100.0%

Warning: these figures indicate the number of cases found in the data file. They cannot be interpreted as summary statistics of the population of interest.

idno: idno

Information	[Type= continuous] [Format=numeric] [Range= 1-30000] [Missing=*]
Statistics [NW/ W]	[Valid=30000 /-] [Invalid=0 /-]
Definition	Person ID number. Numeric IDs long integer format, unique for an individual

hhold_id: hhold_id

Information	[Type= continuous] [Format=numeric] [Range= 10101001-60505010] [Missing=*]
Statistics [NW/ W]	[Valid=30000 /-] [Invalid=0 /-]
Definition	Household ID number. Geographical location, Should be unique for each household

hhold_id_extra: hhold_id_extra

Information	[Type= discrete] [Format=numeric] [Missing=*]
Statistics [NW/ W]	[Valid=0 /-] [Invalid=30000 /-]
Definition	Household ID number. If site has another definition for Households (e.g. social units) include it here

sex: sex

Information	[Type= discrete] [Format=numeric] [Range= 1-2] [Missing=*]
Statistics [NW/ W]	[Valid=30000 /-] [Invalid=0 /-]
Definition	Sex / Gender of study participant. Must not vary between residence episodes

Value	Label	Cases	Percentage
1	Male	14944	49.8%
2	Female	15056	50.2%

Warning: these figures indicate the number of cases found in the data file. They cannot be interpreted as summary statistics of the population of interest.

dob: dob

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=29981 /-]
Definition	Date of birth- best estimate. If actual month and day are not known it is OK to impute, e.g. assign to middle of the month or mid-year. Must not vary between residence episodes

residence: residence

Information	[Type= discrete] [Format=numeric] [Range= 1-3] [Missing=*]
Statistics [NW/ W]	[Valid=30000 /-] [Invalid=0 /-]
Definition	Type of area within DSS . Aim to distinguish urban / rural, or among rural areas distinguish remote / roadside, or by dominant industry

File : synthetic_residencies_v1_0

residence: residence

Value	Label	Cases	Percentage
1	Rural	10057	33.5%
2	Semi-Urban	10001	33.3%
3	Urban	9942	33.1%

Warning: these figures indicate the number of cases found in the data file. They cannot be interpreted as summary statistics of the population of interest.

entry_type: entry_type

Information	[Type= discrete] [Format=numeric] [Range= 1-4] [Missing=*]
Statistics [NW/ W]	[Valid=30000 /-] [Invalid=0 /-]
Definition	A code identifying the event that has occurred. Type of event.

Value	Label	Cases	Percentage
1	Baseline recruitment	7515	25.0%
2	Birth	7413	24.7%
3	External in-migration	7544	25.1%
4	Internal in-migration	7528	25.1%
5	Found after lost to follow up	0	
6	Became eligible for study	0	

Warning: these figures indicate the number of cases found in the data file. They cannot be interpreted as summary statistics of the population of interest.

entry_date: entry_date

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=29996 /-]
Definition	Date on which the event occurred.

entry_type_of_date: entry_type_of_date

Information	[Type= discrete] [Format=numeric] [Range= 1-1] [Missing=*]
Statistics [NW/ W]	[Valid=29996 /-] [Invalid=4 /-]
Definition	Description of how event_date was obtained

Value	Label	Cases	Percentage
1	Reported by HH informant at interview	29996	100.0%
2	Reported by key informant	0	
3	Imputed	0	
Sysmiss		4	

Warning: these figures indicate the number of cases found in the data file. They cannot be interpreted as summary statistics of the population of interest.

entry_obs_date: entry_obs_date

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=29996 /-]
Definition	Observation date. Date on which the event was observed (recorded), also known as surveillance visit date

entry_obs_round: entry_obs_round

Information	[Type= continuous] [Format=numeric] [Range= 1-25] [Missing=*]
Statistics [NW/ W]	[Valid=30000 /-] [Invalid=0 /-]
Definition	Observation round.

File : synthetic_residencies_v1_0

entry_obs_round: entry_obs_round

	Surveillance round when the event was observed (recorded), also know as surveillance round

exit_type: exit_type

Information	[Type= discrete] [Format=numeric] [Range= 11-15] [Missing=*]
Statistics [NW/ W]	[Valid=30000 /-] [Invalid=0 /-]
Definition	A code identifying the event that has occurred. Type of event

Value	Label	Cases	Percentage
11	Present in study site	5938	19.8%
12	Death	6198	20.7%
13	Out-migration	5938	19.8%
14	Internal out-migration	5932	19.8%
15	Lost to follow-up	5994	20.0%
16	Became ineligible for study	0	

Warning: these figures indicate the number of cases found in the data file. They cannot be interpreted as summary statistics of the population of interest.

exit_date: exit_date

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=29984 /-]
Definition	Date on which the event occurred.

exit_type_of_date: exit_type_of_date

Information	[Type= discrete] [Format=numeric] [Range= 1-1] [Missing=*]
Statistics [NW/ W]	[Valid=29984 /-] [Invalid=16 /-]
Definition	Description of how event_date was obtained

Value	Label	Cases	Percentage
1	Reported by HH informant at interview	29984	100.0%
2	Reported by key informant	0	
3	Imputed	0	
Sysmiss		16	

Warning: these figures indicate the number of cases found in the data file. They cannot be interpreted as summary statistics of the population of interest.

exit_obs_date: exit_obs_date

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=30000 /-]
Definition	Observation date. Date on which the event was observed (recorded), also known as surveillance visit date

exit_obs_round: exit_obs_round

Information	[Type= continuous] [Format=numeric] [Range= 1-27] [Missing=*]
Statistics [NW/ W]	[Valid=30000 /-] [Invalid=0 /-]
Definition	Observation round. Surveillance round when the event was observed (recorded), also know as surveillance round

File : synthetic_hiv_testing_v1_0

idno: idno

Information	[Type= continuous] [Format=numeric] [Range= 2-51000] [Missing=*]
Statistics [NW/ W]	[Valid=15000 /-] [Invalid=0 /-]
Definition	Numeric IDs long integer format, unique for an individual

study_name: study_name

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=15000 /-] [Invalid=0 /-]
Definition	Character - consistent across data sets

test_report_date: test_report_date

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=15000 /-]

hiv_test_date: hiv_test_date

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=15000 /-]
Definition	If test carried out in survey or study clinic date will be known exactly, if retrospectively reported by respondent may be approximated to mid-month or mid-year.

hiv_test_result: hiv_test_result

Information	[Type= discrete] [Format=numeric] [Range= 0-8] [Missing=*]
Statistics [NW/ W]	[Valid=11234 /-] [Invalid=3766 /-]
Definition	Indeterminate means test was part of study, but results were inconclusive; not reported means that participant said they had an HIV test outside of study setting but did not disclose result in interview. If participant says they do not know test result, code this as not reported, but only if your research study has no record of the result. If result is recorded in research study or clinic data base do not use not reported code.

Value	Label	Cases	Percentage
0	Negative	3746	33.3%
1	Positive	3752	33.4%
2	Indeterminate	0	
3	Not reported	0	
8		3736	33.3%
Sysmiss		3766	

Warning: these figures indicate the number of cases found in the data file. They cannot be interpreted as summary statistics of the population of interest.

informed_of_result: informed_of_result

Information	[Type= discrete] [Format=numeric] [Range= 0-1] [Missing=*]
Statistics [NW/ W]	[Valid=15000 /-] [Invalid=0 /-]
Definition	No codes typical for anonymised tests in sero-surveys; yes codes typical for VCT or PICT. It is possible for participant to say they were informed of result, even if they did not want to report it in survey interview.

Value	Label	Cases	Percentage
0	No	7552	50.3%
1	Yes	7448	49.7%
8	Don’t know/not asked	0	

Warning: these figures indicate the number of cases found in the data file. They cannot be interpreted as summary statistics of the population of interest.

File : synthetic_hiv_testing_v1_0

source_of_test_information: source_of_test_information

Information	[Type= discrete] [Format=numeric] [Range= 1-1] [Missing=*]
Statistics [NW/ W]	[Valid=15000 /-] [Invalid=0 /-]
Definition	To distinguish between routine tests in surveys, extra tests in study clinics with results directly recorded in study data base, self disclosed results and results from VA proxy respondents

Value	Label	Cases	Percentage
1	Part of a population based study	15000	100.0%
2	Part of a special research study	0	
3	Clinical record- HIV clinic	0	
4	Self-reported by respondent	0	
5	Report by proxy respondent at VA	0	
6	Clinical record- walk in VCT	0	
7	Clinical record- PMTCT/ANC	0	
8	Clinical record- other	0	

Warning: these figures indicate the number of cases found in the data file. They cannot be interpreted as summary statistics of the population of interest.

test_assumption: test_assumption

Information	[Type= discrete] [Format=numeric] [Range= 0-0] [Missing=*]
Statistics [NW/ W]	[Valid=15000 /-] [Invalid=0 /-]
Definition	Most sites will code all tests as “0” for this variable. Sites that do not test, in their population based study, someone who they have recently tested as part of a special study should code this variable as “1”. Those studies that do not test repeat positives should use code “2” here for people who were not tested because they were already tested and found positive in the past. It is essential that these codes are only used on people who participated in the study and did not refuse to test. Use them only for people who were not tested simply due to study protocol, as outlined above.

Value	Label	Cases	Percentage
0	New test	15000	100.0%
1	Test from previous study used	0	
2	Previous positive HIV test used	0	
3	Self reported instead of testing	0	

Warning: these figures indicate the number of cases found in the data file. They cannot be interpreted as summary statistics of the population of interest.

original_hiv_test_result: original_hiv_test_result

Information	[Type= discrete] [Format=numeric] [Missing=*]
Statistics [NW/ W]	[Valid=0 /-] [Invalid=15000 /-]
Definition	This is the HIV test result before any changes were made for example the final test result “hiv_test_result” may be missing but was originally reported as negative. This may arise due to site specific discussions about retro converters and how to deal with them. It should usually be identical to “hiv_test_result”.

Value	Label	Cases	Percentage
0	Negative	0	
1	Positive	0	
2	Indeterminate	0	
3	Not reported	0	
Sysmiss		15000	

Warning: these figures indicate the number of cases found in the data file. They cannot be interpreted as summary statistics of the population of interest.

survey_round_name: survey_round_name

Information	[Type= discrete] [Format=character] [Missing=*]

File : synthetic_hiv_testing_v1_0

survey_round_name: survey_round_name

Statistics [NW/ W]	[Valid=15000 /-] [Invalid=0 /-]
Definition	If the testing was done as part of a survey this variable should denote the name of the survey for example “sero1” or “TBSurvey2” etc it should be in string format.

Value	Label	Cases	Percentage
Sero 0		380	2.5%
Sero 1		1328	8.9%
Sero 2		1763	11.8%
Sero 3		1763	11.8%
Sero 4		1430	9.5%
Sero 5		1912	12.7%
Sero 6		2113	14.1%
Sero 7		2854	19.0%
Sero 8		1457	9.7%

Warning: these figures indicate the number of cases found in the data file. They cannot be interpreted as summary statistics of the population of interest.