Malawi IDSR COVID-19 Synthetic DataSet

Version v1.0

Date: 26/01/2022

Documentation Control Sheet

It may be necessary to issue amendments or clarifications to parts of this document. This form must be updated whenever changes are made and should be filled inside the front cover of the new or amended document.

Version	Summary of Change	Prepared By	Date	Reviewed By	Date
1.0		Amelia Taylor and Tuntufye Mwanyongo	26/01/2022		

Contact: ataylor@mubas.ac.mw

About the Dataset

This dataset is based on the IDSR Case Based Reporting Form used to record COVID-19 cases and tests by the Ministry of Health in Malawi. The form can be found in the archive for the dataset. This form is based on the general IDSR Case Based form in use in Malawi by adding COVID-19 specific fields from the WHO CRF for COVID-19. When developing this dataset we did not have access to the real data collected in Malawi. Hence this data does not contain any sensitive information.

The development of the dataset was done at the Malawi University of Business and Applied Sciences to be used for the INSPIRE PEACH project. INSPIRE PEACH aims at building a data hub for standardizing and harmonizing Covid-19 data from different sources in Malawi and Kenya using the OMOP CDM. A challenge that INSPIRE PEACH faced both in Malawi and Kenya was access to COVID-19 data: protocols to access data are time consuming, not all the data is available in digital form and there are cross border sharing restrictions. Therefore, it looked practical to generate synthetic data from a small subset of the real world data. The aim of the dataset was to be used in testing the data pipeline for the INSPIRE PEACH project and also to showcase to various stakeholders the process of loading data to OMOP.

The data generated in this version 1.0 is based on purely random generated values and takes into account individual level fidelity metrics. It does not take into account metrics that account for population fidelity of the data.

Count

Total records/cases: 10000

Cases with negative lab result: 4986

Cases with positive lab result: 5014

Dataset

Synthetic data was generated using python code and saved to a csv file called Malawi_IDSR_Covid19SyntheticDataSet_v10.csv. The fields are separated by “comma”.

Data Description

The data follows the format of the IDSR CASE BASED SURVEILLANCE REPORTING FORM. The fields are listed below alongside their description, domain values, data types and ranges. There are 51 variables in total in the questionnaire. The categories used for each variable are contained in the file Categories.csv.

No.	Variable Name	Field Name	DESCRIPTION	DOMAIN OF VALUES	RANGES	DATA TYPE / FORMAT
1	REPR_FACI	Reporting facility	Names of main health facilities in Malawi[1].	facilities	90 categories	Numeric
2	REPR_DIST	Reporting district	Names of all 28 districts in Malawi[2]	districts	28 categories	Text / ISO 3166-2 codes
3	TYP_CASE	Type of case	Type of hospital visit.	Type of case	3 categories	Text
4	REPR_NM	Reporter name	Names of the person who filled the form.	List of ids	MOH1 - MOH200	Text
5	REPR_PHON_NUM	Reporting Phone Number	Phone numbers of reporters	Phone numbers		Character
6	REPR_D	Reporting date	The date when the current incident/ case is reported.	Faker List of dates		Date (YYYY-MM-DD)
7	TYP_REPR_DISE_COND	Type of reporting disease condition	Disease names related to COVID-19, i.e. SARS-COV-2, AEFI,SARI	Disease conditions	3 categories	Text
8	LN_CASE	Last Name of case	Surnames.	Names of people.		Text
9	FN_CASE	First Name of case	First names.	Names of people.		Text

[1] Source: Baobab Health Trust List of Facilities https://github.com/BaobabHealthTrust/master-facility-list/blob/master/health-facilities.csv We selected only District Hospitals, Central Hospitals, Rural/Community Hospitals and Govt/Public in 28 districts.
[2] Source: https://en.wikipedia.org/wiki/ISO_3166-2:MW

10	DOB[3]	Date of birth	Dates of birth.	Dates of birth	Min: 1920/1/1 Max: 2000/12/31	Date[4]
11	NATL	Nationalities	Common nationalities who reside in or frequently travel in and out of Malawi[5].	Nationalities	33 categories	Text
12	CASE_UID1	Cases UID (1)	Type of identification document.	Cases_UID	2 Categories	Character
13	CASE_UID2	Case UID (2)	Personal identification details, e.g. ID or passport numbers.		2 categories	Numeric
14	DIST_CASE_RESI	District of case residence	District in Malawi where the case resides.	District of case residence		Text
15	DIST_CASE_RESI_TYP	District of case residence (Type)	Type of residence in the district	Residence type	2 categories	Text
16	SEX	Sex	Sex	Sex	2 categories Male = 1, Female = 2	Text
17	OCCU	Occupation	Job name of patient or case.	Faker List of category of occupations	17 categories	Text
18	PHYS_ADD	Physical	Address	Faker List		Characte

[3] The IDSR form has an extra data element, which is Age that can be completed when DOB is not available. For simplicity we generate only DOBs and then ages can be derived from these.
[4] All dates are in the YYYY-MM-DD format.
[5] We included all countries in which Malawi has diplomatic missions and other nationalities known to frequently visit Malawi (Source: https://www.malawitourism.com/visa-guide/).

	R	address		of Physical addresses		r
19	NEAR_LMK	Nearest Landmark	Type of nearest publicly recognizable landmark	Nearest_Landmark		Text
20	PHON_NUM_CASE	Phone number of case	Phone numbers	Faker List of phone numbers		Char
21	PARE_CARET_NM	Parent or Caretaker name	Name of Parent or caretaker	Faker List of People's names		Text
22	D_SEEN_FACI	Date seen at facility	Date of first identification of the case.	Faker List of dates	Min: 2020/1/1 Max: 2021/12/31	Date
23	VACC	Vaccination	Type of vaccination	vaccination	2 categories	Text
24	VACC_TYP_VAC	Vaccination type of vaccine	Name of covid-19 vaccine.	Vaccination_type_of_vaccine	2 categories	Text
25	VACC_NUM_DOSE	Vaccination (# of doses)	Number of covid-19 vaccine doses received by a case.	Vaccination (# of doses)	Min: 0 Max: 5	Integer
26	D_FACI_NOTI_DIST	Date facility notified district	Date when a facility notified the district of the results of a lab test.	Faker List of dates	Min: 2020/1/1 Max: 2021/1/1	Date (YYYY-MM-DD)
27	D_L_VACCI	Date of last vaccination	Date when the case got the last vaccination mentioned.	Faker List of dates	Min: 2021/1/1 Max: 2021/12/31	Date (YYYY-MM-DD)
28	RECE_TRAV_HIST	Recent travel history	The type of recent travel.	Recent_travel_history	2 categories	Text

29	D_O_RETU	Date of return	Date when the case returned home after traveling.	Faker List of dates	Min: 2020/1/1 Max: 2021/12/31	Date
30	ANY_CONT_OT_CASE	Any contact with OT case	Type of contact with someone who is suspected or confirmed to be covid19 positive.	Any_contact_with_OT_case	3 categories	Text
31	ANY_CLUSTER	Any clustering	Type of clustering.	Any_clustering	5 categories	Text
32	D_ONSE	Date of onset	Date of onset of the symptoms mentioned on the form.	Faker List of dates	Min: 2020/1/1 Max: 2021/3/31	Date
33	PREG_CASE_FEM	Pregnancy (If case is female)	Whether the case is pregnant or not.	Pregnancy	2 Categories	Text
34	TRIM	Trimester	The trimester of the pregnancy (if case is pregnant)	Trimester	3 Categories	Text
35	PRES_SYMP	Presenting symptoms	A list of symptoms from a predefined list.	Presenting Symptoms	14 categories.	Text
36	U_CONDS	Underlying conditions	Underlying conditions taken from a predefined list defined by the MOH as being relevant.	Underlying conditions	15 categories	Text
37	COVI_HIST	Covid(+)Hist	If case has any covid positive history (Previously covid infected)	Covid(+)Hist		Text

38	P_COMPL_FRM_NM	Personal Completer Form(Name)		Faker List of names of people		Text
39	D_S_COLL	Date Specimen collected	Date specimen was collected from case	Faker List of dates	Min: 2020/1/1 Max:2021/6/31	Date (YYYY-MM-DD)
40	D_S_SENT_LAB	Date Specimen sent to lab	Date specimen was sent to lab	Faker List of dates	Min: 2021/1/1 Max: 2021/12/31	Date (YYYY-MM-DD)
41	S_TYP	Specimen Type	Type of specimen	Specimen Type	4 categories	Text
42	S_COND	Specimen Condition	Condition of the specimen	Specimen_ Condition	2 categories	Text
43	D_LAB_RECE_S	Date lab received specimen	Date lab received specimen	Faker List of dates	Min: 2021/5/1 Max:2021/6/31	Date (YYYY-MM-DD)
44	TYP_O_TESTS_PERF	Type of Tests Performed	Type of covid test	Type of Tests Performed	2 categories	Text
45	TEST_PLAT	Testing platform	Type of covid testing platform	Testing platform	2 categories	Text
46	FIN_LAB_RESU	Final Laboratory Result	The result of the test.	Final_Laboratory_Result	2 categories	Text
47	D_LAB_SENT_RESU_DIST	Date Lab sent Result to District	Date lab sent result to district	Faker List of dates	Min: 2021/6/1 Max:2021/8/31	Date (YYYY-MM-DD)
48	D_RESU_SENT_HCW	Date Result sent to HCW	Date result sent to HCW	Faker List of	Min:2021/8/1	Date (YYYY-

				dates	Max: 2021/9/31	MM-DD)
49	D_DIST_RECE_RESU	Date district received result	Date district received result	Faker List of dates	Min: 2021/9/1 Max: 2021/12/31	Date (YYYY-MM-DD)
50	CASE_F_O	Case Final Outcome	Case final outcome	Case_final_ outcome	4 categories	Text
51	CASE_F_CLASS	Case Final Classification	Case final classification	Case final classificatio n	4 categories	Text

Methodology

We use Python libraries to generate values for the 51 variables. To make the data as realistic as possible, given that we do not have access to a dataset of real data, we set up a database of categories (options) to be used for some of the variables. These options are saved in an Categories.csv file in the archive for the dataset. These will be used by the Python generator to make sure that certain individual level constraints are satisfied.

The data we generate satisfies the following individual level data fidelity metrics:

1. **Names and Gender Correspondence**: for example typical male names correspond to Male (sex). We generated only first names and codes for Surnames to avoid a situation where the surnames have been all European.

2. **Facilities and Districts are linked:** a facility is expected to be located in a certain district.

3. **Linked and Staggered Dates:** The following order/priority of dates was used. For example, lab dates should be in the expected order, e.g. date when the specimen was sent to the lab cannot be after the date in which the result is sent back to the health center (see the table above with the order in which dates were generated).

Time Interval between Dates	0 .. 20		0..5	0..1	0..3	0...5	0...3
Order of Generation of Dates	2	1	3	4	5	6	7
	D_L_VACCI	D_SEEN_FACI	D_REPORTING				
	D_O_RETU		D_FACI_NOTI_DIST				
	D_ONSE		D_S_COLL	D_S_SENT_LAB	D_LAB_RECE_S	D_LAB_SENT_RESU_DIST	D_DIST_RECE_RESU
						D_RESU_SENT_HCW	

4. **Date of Birth should be in the past:** e.g. 1920 - 2000.

5. **Pregnancy**: All males have a null response while females have a yes or no response in the Pregnancy field.

6. **Trimester** : All males have a null response while pregnant females have a value in the Trimester field.

7. **Addresses:** we generated addresses that are more realistic for the Malawi context, P.O.Box [No] District.

8. **Phone Numbers**: we also generated phone Malawian numbers which start with the prefix +265.

9. **Occupations**: we tried to limit these values to typical occupations in Malawi.

Tools

Data for some of the columns like name, occupation, address and date (all datetime related columns) were generated using faker python library[6] while the rest of the columns were generated using random python module[7]. The implication is that all the data generated with Faker uses values from a standard database, hence these are not specific to a Malawian context, for example names of people and addresses. As these elements will most likely be later on de-identified, the use of Faker was deemed sufficient. For most of the variables that need to satisfy some individual level fidelity measures, we used the python random module in connection with the categories in the Categories.csv.

LIMITATIONS

We do not take into account data fidelity metrics at the population level. For example, the age split of our dataset does not correspond to realistic values. The same for other variables, e.g. case outcomes. For variables where we use the random python module, the values are equally

[6] https://faker.readthedocs.io/en/master/
[7] https://docs.python.org/3/library/random.html

distributed according to the number of categories for that field. For example, we have 2 types of outcomes for the laboratory results, hence the positive cases are almost equal to the negative cases.

IMPROVEMENTS

Version	Improvement
1.1	Incorporate data fidelity measures at the population level: (1) demographics such as age and sex; (2) COVID - 19 specific statistics: number of positive tests, negative tests to match the country statistics; testing numbers over periods of time to be realistic; etc.

USAGE

This dataset is purely synthetic and should not be used in publishing results out of it. It can be used for testing and simulation purposes only. No real world data has been collected for this dataset.

Appendix A

Correspondence between fields in the questionnaire and variables.

Questionnaire Field Names	Variables
Reporting Facility	REPR_FACI
Reporting District	REPR_DIST
Type of Case	TYP_CASE
Reporter Name	REPR_NM
Reporting Phone #	REPR_PHON_NUM
Reporting Date	REPR_D

Questionnaire Field Names	Variables
Type of Reporting Disease-Condition	TYP_REPR_DISE_COND
Last Name of Case	LN_CASE
First Name of Case	FN_CASE
Date of Birth (dd/mm/yyyy)	DOB
Nationality	NATL
Cases UID	CASE_UID1
Cases UID2	CASE_UID2
District of Case Residence	DIST_CASE_RESI
District of Case Residence(Type)	DIST_CASE_RESI_TYP
Sex	SEX
Occupation	OCCU
Physical Address	PHYS_ADDR
Nearest Landmark	NEAR_LMK
Phone number of Case	PHON_NUM_CASE
Parent or Care Taker Name	PARE_CARET_NM
Date Seen at Facility	D_SEEN_FACI
Vaccination	VACC
Vaccination_type_of vaccine	VACC_TYP_VAC
Vaccination(# of doses)	VACC_NUM_DOSE
Date Facility Notified District	D_FACI_NOTI_DIST
Date of Last Vaccination	D_L_VACCI
Recent Travel History	RECE_TRAV_HIST
Date of Return	D_O_RETU
Any contact with OT case	ANY_CONT_OT_CASE
Any Clustering	ANY_CLUSTER

Questionnaire Field Names	Variables
Date of Onset	D_ONSE
Pregnancy (if case is female)	PREG_CASE_FEM
Trimester	TRIM
Presenting Symptom(s), tick if any presented	PRES_SYMP
Underlying Condition(s), tick if any presented	U_CONDS
Covid(+)Hist	COVI_HIST
Person Completer Form(Name)	P_COMPL_FRM_NM
Date specimen collected	D_S_COLL
Date specimen sent to lab	D_S_SENT_LAB
Specimen type	S_TYP
Specimen condition	S_COND
Date lab received specimen	D_LAB_RECE_S
Type of test(s) performed	TYP_O_TESTS_PERF
Testing Platform	TEST_PLAT
Final Laboratory Result	FIN_LAB_RESU
Date lab sent result to district	D_LAB_SENT_RESU_DIST
Date result sent to HCW	D_RESU_SENT_HCW
Date district received result	D_DIST_RECE_RESU
Case Final Outcome	CASE_F_O
Case Final Classification	CASE_F_CLASS

_____END OF DOCUMENT _____