

SAPRIN Mental Health Data Prize 2022

Prof Mark Collinson, Dr Kobus Herbst, Prof Steve Tollman, Prof Eric Maimela, Prof Willem Hanekom

Report generated on: August 7, 2022

Visit our data catalog at: <https://saprindata.samrc.ac.za/index.php>

Overview

Identification

ID NUMBER

SAPRIN.SMHDP2022V1

Version

VERSION DESCRIPTION

v1: Dataset for public distribution.

PRODUCTION DATE

2022-08-01

NOTES

v1: Dataset for public distribution.

Overview

ABSTRACT

SAPRIN (South African Population Research Infrastructure Network) is a network of health and demographic surveillance sites in South Africa that consists of five Health and Demographic Surveillance System (HDSS) nodes located in South Africa.

Between them, the nodes follow more than 75 000 households (320 000 individuals) longitudinally through regular surveillance visits. Shortly after the start of the Covid-19 pandemic, SAPRIN implemented a shared Covid-19 surveillance programme in the MRC/Wits University Agincourt HDSS in Bushbuckridge District, Mpumalanga, established in 1993; the University of Limpopo DIMAMO HDSS in the Capricorn District of Limpopo, established in 1996, and the Africa Health Research Institute (AHRI) HDSS in uMkhanyakude District, KwaZulu-Natal, established in 2000. This Covid-19 surveillance is still being conducted in these three SAPRIN nodes.

As part of the Covid-19 surveillance, the PHQ-2 and GAD-2 screening questions were administered to household respondents in Agincourt, DIMAMO and AHRI. By the end of 2021, a total of 90 000 such interviews were conducted, with approximately 12 000 interviews in the target group of 14-24-year-olds. Although the PHQ-2 and GAD-2 questions on their own are not likely to be of interest to the Mental Data prize participants, several factors make this dataset to be of greater interest:

The interviews can be directly linked to the detailed longitudinal surveillance data in the three nodes, providing interesting contextual data to this set of observations on depression and anxiety during the span of the covid epidemic from its start across several epidemic waves of infection in South Africa.

These contextual data include:

- Individual-level data on partnership status, educational attainment, and employment, including self-reported health status.
- Household data of the households the respondents are members of, including household composition and whether the respondents' parents are co-resident with them.
- Household socio-economic status.
- Household asset status
- A large set of Covid-19 specific data, including vaccine acceptance and hesitancy data and the impact of Covid-19 measures on the household.

The Covid-19 specific interviews were conducted from May 2020 and are still ongoing, with more than one interview with some participants at different points in time, allowing for the analysis of temporal effects.

The SAPRIN Mental Health Data Prize 2022 datasets consists of six types of the Demographic surveillance datasets :

1. SAPRIN Individual exposure episodes. This dataset splits the basic surveillance episodes at calendar year-end and at the date when the age in years (birthday) of an individual change. In the case of women who have given births, episodes are split at the time of delivery as well.
2. SAPRIN Individual status observations. This dataset consists of status observations such as education, employment, employment and partnership status of an individual, that recur at more or less regular interval per individual over the study period.
3. SAPRIN household status observations. This dataset consists of socio-economic status observations for a household. This data is collected from a household proxy respondent, preferably the head of household or any next available senior adult resident household member at more or less regular interval over the duration of the study.
4. SAPRIN household asset status observations. This dataset consists of asset status observations for a household. This data is collected from a household proxy respondent, preferably the head of household or any next available senior adult resident household member at more or less regular interval over the duration of the study.
5. SAPRIN individual COVID-19. This dataset consists of Covid-related status observations pertaining to Covid-19 diagnosis, vaccination status, attitudes to vaccination and the PHQ-2 and GAD-2 mental health related questions.
6. SAPRIN household COVID-19. This dataset consists of Covid-19 related household level status observations, household awareness, and impact of Covid-19 control measures on the household.

KIND OF DATA

Event history data

UNITS OF ANALYSIS

Individual and household interviews

Scope

NOTES

Each record in the exposure dataset represents a period of observation for an individual during which all the recorded characteristics of the individual stay constant. For example, on the birthday of the individual a new episode will start, because the age of the individual has changed. Any change in one of the status values, such as education or marital status, will likewise result in a new episode on the date of the change. For the COVID-19 data, the questionnaire included both household-level and individual-specific questions, the latter of which could be directly addressed by other household members if they were present. The primary respondent acted as a proxy in all other cases. COVID-19 symptom screening was included in the questionnaire.

TOPICS

Topic	Vocabulary	URI
Mental Health, Covid-19		

KEYWORDS

Mental Health, Covid-19

Coverage

GEOGRAPHIC COVERAGE

SAPRIN (South African Population Research Infrastructure Network) is a network of health and demographic surveillance sites in South Africa that consists of five Health and Demographic Surveillance System (HDSS) nodes located in South Africa, namely: 1) MRC/Wits University Agincourt HDSS in Bushbuckridge District, Mpumalanga, which has collected data since 1993. The nodal website is <http://www.agincourt.co.za>. 2) the University of Limpopo DIMAMO HDSS in the Capricorn District of Limpopo, which has collected data since 1996. The nodal website is: N/A. 3) the Africa Health Research Institute (AHRI)

HDSS in uMkhanyakude District, KwaZulu-Natal, which has collected data since 2000. The nodal website is <http://www.ahri.org>. 4) the Gauteng Research Triangle Initiative for the Study of Population, Infrastructure and Regional Economic Development (GRT-INSPIRED) in Hillbrow, Johannesburg, and Atteridgeville and Melusi, Tshwane, Gauteng. The nodal website is: N/A. 5) and the Cape Town Surveillance through Healthcare Action Research Project (C-SHARP), Nomzamo and Bishop Lavis, Cape Town, Western Cape. The nodal website is: N/A.

UNIVERSE

Eligibility covered all individuals who are between the ages of 14 and 24yrs on the 1st of January 2020 and resident within each of the three SAPRIN nodal sites. Residence was defined as intention to sleep the majority of time at the dwelling in these areas over a four-month period

Producers and Sponsors

PRIMARY INVESTIGATOR(S)

Name	Affiliation
Prof Mark Collinson	SAPRIN
Dr Kobus Herbst	SAPRIN
Prof Steve Tollman	Agincourt
Prof Eric Maimela	DIMAMO
Prof Willem Hanekom	AHRI

OTHER PRODUCER(S)

Name	Affiliation	Role
Molulaqhoora Linda Maoyi	SAPRIN	Technical Assistance
Tinofa Mutevedzi	SAPRIN	Technical Assistance
Chodziwadziwa Kabudula	Agincourt	Technical Assistance
Joseph Tlouyamma	DIMAMO	Technical Assistance
Dickman Gareta	AHRI	Technical Assistance

FUNDING

Name	Abbreviation	Role
Department of Science and Innovation	DSI	Current Funder

OTHER ACKNOWLEDGEMENTS

Name	Affiliation	Role
Agincourt Data Team	Agincourt	Providing Data
DIMAMO Data Team	DIMAMO	Providing Data
AHRI Data Team	AHRI	Providing Data
Steve Tollman	Agincourt	
Eric Maimela	DIMAMO	
Willem Hanekom	AHRI	
Centre for High Performance Computing	Centre for High Performance Computing	Providing IT Infrastructure for Data Processing

Metadata Production

METADATA PRODUCED BY

Name	Abbreviation	Affiliation	Role
Molulaqhooa Linda Maoyi	MLM	SAPRIN	Documentation of Study and Review of the metadata
Kobus Herbst	KH	SAPRIN	Documentation of Study and Review of the metadata
Tinofa Mutevedzi	TM	SAPRIN	Documentation of Study and Review of the metadata
Mark Collinson	MC	SAPRIN	Documentation of Study and Review of the metadata

DATE OF METADATA PRODUCTION

2022-08-07

DDI DOCUMENT VERSION

Version 2 (August 2022)

DDI DOCUMENT ID

DDI.SAPRIN.SMHDP2022V1

Sampling

Sampling Procedure

All individuals who are between the ages of 14 and 24yrs on the 1st of January 2020. All exposure episodes of these individuals from the start of exposure to the 30th of April 2022 are included in the dataset.

Questionnaires

Overview

The data on this Repository is not the result of a single questionnaire but is a result of harmonised data from three different sites longitudinally collected over more than twenty years using different questionnaires that varied over time and site.

Data Collection

Data Collection Dates

Start	End	Cycle
1993-01-01	2022-04-30	Agincourt
1996-01-01	2022-04-30	DIMAMO
2000-01-01	2022-04-30	AHRI

Time Periods

Start	End	Cycle
1993-01-01		Agincourt
1996-01-01		DIMAMO
2000-01-01		AHRI

Data Collection Notes

In all the HDSS nodes, data are collected from a household proxy respondent, preferably the head of household or any next available senior adult resident household member, after informed consent was obtained by trained fieldworkers. Respondents are informed of the purpose and confidentiality of the interview, their right to refuse participation or withdraw from the study, and that scientists would be given access to anonymised data to analyse and publish information. Informed consent was verbal in all HDSS nodes until 2016. Written informed consent started in 2017 in AHRI, and 2018 in DIMAMO and 2019 in Agincourt. Until 2016 for Agincourt and AHRI, and 2017 for DIMAMO, data collection was field-based 'paper and pen' personal interviews (PAPI), before changing to field-based computer-assisted personal interviews (CAPI). Since 2019, all SAPRIN HDSS nodes collect data in 3 annual rounds over a 45-week data collection schedule; one field-based CAPI round, sandwiched on either side by a Call-Centre-based computer assisted telephonic interview (CATI), to create 3 data points at an interval of approximately 4 months in each calendar year. In the past HDSS nodes had different data collection frequencies. AHRI data collection was 2 PAPI rounds per year from inception to 2011, changing to 3 PAPI rounds per year between 2012 and 2016, before becoming 1 PAPI round and 2 CATI rounds from 2017. Agincourt and DIMAMO have been collecting data once annually in a census-type format, over 4-5-month period until 2018.

Questionnaires

The data on this Repository is not the result of a single questionnaire but is a result of harmonised data from three different sites longitudinally collected over more than twenty years using different questionnaires that varied over time and site.

Data Processing

Data Editing

The first step in the data preparation process is quality assurance. The SAPRIN Management hub team assess the data submitted to ensure it is in the correct format and falls within expected value ranges. Other potential issues checked include: missing data, incorrect data types, unexpected duplicate or orphan records. The SAPRIN Management hub assess this conversion by running both original operational database and the SAPRIN database created from the operational database through the iSHARE data quality assessment and indicator process. The data quality checking process is conducted using Pentaho Data Integration (PDI). PDI provides the Extract, Transform, and Load (ETL) capabilities that facilitates the process of capturing, cleansing, and storing data using a uniform and consistent format that is accessible and relevant to end users. The principle of the data quality checks is that if the data conversion conducted by the nodes was complete and accurate, there should be little or no difference in the data quality and demographic indicators between the base and SAPRIN versions of the nodal data. If the data submitted by the nodes meets the criteria for inclusion into the consolidated dataset the data moves to the second step of the data production process. However, if the data fail the inclusion checks, this could then lead to another iteration of data submission and quality control checks until SAPRIN Management hub is satisfied that they have high quality data. To produce this final standard dataset, the data is processed using PDI on the Centre for High Performance Computing cluster.

Data Appraisal

Estimates of Sampling Error

Not Applicable

File Description

Variable List

SAPRIN.MHDPIEE2022

Content

Cases 1126280

Variable(s) 24

Structure Type:
Keys: ()

Version

Producer

Missing Data

Variables

ID	NAME	LABEL	TYPE	FORMAT	QUESTION
V1	NodeId	SAPRIN Node identifier	discrete	numeric	
V2	IndividualId	Unique individual identifier	contin	numeric	
V3	DoB	Date of birth	discrete	character	
V4	DoD	Date of death	discrete	character	
V5	CalendarYear	The calendar year in which the episode false	contin	numeric	
V6	Age	The age in completed years of the individual	contin	numeric	
V7	Sex	Sex of individual	discrete	numeric	
V8	LocationId	Where individual was resident, household residence if non-resident	contin	numeric	
V9	HouseholdId	Unique household identifier of the household the individual is a member of	contin	numeric	
V10	HHRelation	Relationship to head of household at start of episode	discrete	numeric	
V11	IsUrbanOrRural	Settlement pattern at location	discrete	numeric	
V12	MotherId	Mother's IndividualId	contin	numeric	
V13	FatherId	Father's IndividualId	contin	numeric	
V14	SpouseId	IndividualId of spouse - not available	discrete	numeric	
V15	StartDate	Start date of episode (inclusive)	discrete	character	
V16	EndDate	End date of episode (inclusive)	discrete	character	
V17	StartType	What triggered this episode start?	discrete	numeric	
V18	EndType	How did this episode end?	discrete	numeric	
V19	Episode	This episode number (first=1, last=Episodes)	contin	numeric	
V20	Episodes	Total number of episodes for individual	contin	numeric	
V21	Resident	Whether individual is resident for duration of episode	discrete	numeric	
V22	MotherStatus	Status of mother for duration of episode	discrete	numeric	
V23	FatherStatus	Status of father for duration of episode	discrete	numeric	
V24	ChildrenEverBorn	Number of children ever born to individual	discrete	numeric	

SAPRIN.MHDPISO2022

Content

Cases 8969822

Variable(s) 7

Structure Type:
Keys: ()

Version

Producer

Missing Data

Variables

ID	NAME	LABEL	TYPE	FORMAT	QUESTION
V25	nodeid	Saprin Node identifier	discrete	numeric	
V26	IndividualId	Unique individual identifier (anonymised)	contin	numeric	
V27	ObservationDate	Date of status observation	discrete	character	
V28	SchoolStatus	Education status during this period	discrete	numeric	
V29	PartnershipStatus	Partnership status during this period	discrete	numeric	
V30	HealthStatus	Health status during this period	discrete	numeric	
V31	EmploymentStatus	Employment status during this period	discrete	numeric	

SAPRIN.MHDPHHS2022

Content

Cases 391182

Variable(s) 14

Structure Type:
Keys: ()

Version

Producer

Missing Data

Variables

ID	NAME	LABEL	TYPE	FORMAT	QUESTION
V32	nodeid	Saprin Node identifier	discrete	numeric	
V33	proxyid	Household proxy respondent identifier (anonymised) linkable to IndividualID in e	contin	numeric	
V34	dob	Date of Birth of household proxy respondent	discrete	character	
V35	sex	Sex of household proxy respondent	discrete	numeric	
V36	householdid	Household Id linkable to householdId in episode data	contin	numeric	
V37	observationdate	Date of status observation	discrete	character	
V38	WaterSource	The most commonly used source of drinking water	discrete	numeric	
V39	Toilet	What kind of toilet does the household use?	discrete	numeric	
V40	ElectricitySupply	Is the household connected to the electricity grid?	discrete	numeric	
V41	CookingFuel	What is the main fuel used for cooking?	discrete	numeric	
V42	WallMaterial	What are the construction materials of the walls?	discrete	numeric	
V43	FloorMaterial	What are the construction materials of the floor?	discrete	numeric	
V44	Crime	Has any resident member of the household been a victim of any of these crimes in	discrete	numeric	
V45	FinancialStatus	How would the household classify its financial situation these days ?	discrete	numeric	

SAPRIN.MHDPHHAS2022

Content

Cases 4535997

Variable(s) 5

Structure Type:
Keys: ()

Version

Producer

Missing Data

Variables

ID	NAME	LABEL	TYPE	FORMAT	QUESTION
V46	nodeid	Saprin Node identifier	discrete	numeric	
V47	HouseholdId	Household Id linkable to householdId in episode data	contin	numeric	
V48	ObservationDate	Date of status observation	discrete	character	
V49	AssetId	Asset kind	discrete	numeric	
V50	AssetStatusId	Asset status	discrete	numeric	

SAPRIN.MHDPICS2022

Content

Cases 41509

Variable(s) 52

Structure Type:
Keys: ()

Version

Producer

Missing Data

Variables

ID	NAME	LABEL	TYPE	FORMAT	QUESTION
V51	NodeId	Saprin Node identifier	discrete	numeric	
V52	IndividualId	Unique individual identifier (anonymised)	contin	numeric	
V53	ObservationDate	Date of status observation	discrete	character	
V54	DateOfBirth	Date of Birth of Individual	discrete	character	
V55	PHQ2_1	PHQ2_1	discrete	numeric	
V56	PHQ2_2	PHQ2_2	discrete	numeric	
V57	GAD2_1	GAD2_1	discrete	numeric	
V58	GAD2_2	GAD2_2	discrete	numeric	
V59	rk12	rk12	discrete	numeric	
V60	rk13	rk13	discrete	numeric	
V61	rk11	rk11	discrete	numeric	
V62	xx13	xx13	discrete	numeric	
V63	xx15	xx15	discrete	numeric	
V64	xx16	xx16	discrete	numeric	
V65	xx17	xx17	discrete	numeric	
V66	prior_covid_test	prior_covid_test	discrete	numeric	
V67	prior_covid_dx	prior_covid_dx	discrete	numeric	
V68	prior_covid_dx_date	prior_covid_dx_date	discrete	character	
V69	prior_covid_hosp	prior_covid_hosp	discrete	numeric	
V70	vaccinated	vaccinated_covid	discrete	numeric	
V71	vaccinated_dose2	dose2_vaccinated	discrete	numeric	
V72	vaccinated_dose1_when	dose1_date	discrete	character	
V73	vaccinated_dose2_when	dose2_date	discrete	character	
V74	va12_1	va12_1	discrete	numeric	
V75	va12_2	va12_2	discrete	numeric	
V76	va12_3	va12_3	discrete	numeric	
V77	va12_4	va12_4	discrete	numeric	
V78	va12_5	va12_5	discrete	numeric	

V79	va12_6	va12_6	discrete	numeric
V80	va12_7	va12_7	discrete	numeric
V81	va12_8	va12_8	discrete	numeric
V82	va12_9	va12_none	discrete	numeric
V83	va12_96	va12_other	discrete	numeric
V84	va14_1	va14_1	discrete	numeric
V85	va14_10	va14_10	discrete	numeric
V86	va14_11	va14_11	discrete	numeric
V87	va14_12	va14_12	discrete	numeric
V88	va14_13	va14_13	discrete	numeric
V89	va14_2	va14_2	discrete	numeric
V90	va14_3	va14_3	discrete	numeric
V91	va14_4	va14_4	discrete	numeric
V92	va14_5	va14_5	discrete	numeric
V93	va14_6	va14_6	discrete	numeric
V94	va14_7	va14_7	discrete	numeric
V95	va14_8	va14_8	discrete	numeric
V96	va14_9	va14_9	discrete	numeric
V97	va14_96	va14_none	discrete	numeric
V98	va12_99	va12__99	discrete	numeric
V99	va12_100	va12__100	discrete	numeric
V100	va14_99	va14__99	discrete	numeric
V101	va14_100	va14__100	discrete	numeric
V102	va14_14	va14_14	discrete	numeric

SAPRIN.MHDPHHCS2022

Content

Cases 61397

Variable(s) 84

Structure Type:
Keys: ()

Version

Producer

Missing Data

Variables

ID	NAME	LABEL	TYPE	FORMAT	QUESTION
V103	nodeid	Saprin Node identifier	discrete	numeric	
V104	source	Source of Household Data (F= Field, C= Call Centre)	discrete	character	
V105	proxyid	Unique proxy household identifier (anonymised)	contin	numeric	
V106	householdid	Unique household identifier (anonymised)	contin	numeric	
V107	sex	sex of proxy respondent	discrete	numeric	
V108	dob	Date of Birth of proxy respondent	discrete	character	
V109	observationdate	Date of status observation	discrete	character	
V110	ar01	ar01	discrete	numeric	
V111	ar03__1	ar03__1	contin	numeric	
V112	ar03__2	ar03__2	contin	numeric	
V113	ar03__3	ar03__3	contin	numeric	
V114	ar03__4	ar03__4	contin	numeric	
V115	ar03__5	ar03__5	contin	numeric	
V116	ar03__6	ar03__6	contin	numeric	
V117	ar03__7	ar03__7	contin	numeric	
V118	ar03__8	ar03__8	contin	numeric	
V119	ar03__9	ar03__9	contin	numeric	
V120	ar03__10	ar03__10	contin	numeric	
V121	ar03__11	ar03__11	contin	numeric	
V122	ar03__14	ar03__14	contin	numeric	
V123	ar03__15	ar03__15	contin	numeric	
V124	ar03__16	ar03__16	contin	numeric	
V125	ar03__17	ar03__17	contin	numeric	
V126	ar03__18	ar03__18	contin	numeric	
V127	ar03__19	ar03__19	contin	numeric	
V128	hv02	hv02	contin	numeric	
V129	hi11__1	hi11__1	discrete	numeric	
V130	hi11__2	hi11__2	discrete	numeric	

V131	hi11_3	hi11_3	discrete	numeric
V132	hi11_4	hi11_4	discrete	numeric
V133	hi11_5	hi11_5	discrete	numeric
V134	hi12_1	hi12_1	contin	numeric
V135	hi12_2	hi12_2	contin	numeric
V136	hi12_3	hi12_3	contin	numeric
V137	hi12_4	hi12_4	contin	numeric
V138	hi12_5	hi12_5	contin	numeric
V139	hi12_6	hi12_6	contin	numeric
V140	hi12_7	hi12_7	contin	numeric
V141	xx01	xx01	discrete	numeric
V142	xx11_1	xx11_1	contin	numeric
V143	xx11_2	xx11_2	contin	numeric
V144	xx11_3	xx11_3	contin	numeric
V145	xx11_4	xx11_4	contin	numeric
V146	xx11_5	xx11_5	contin	numeric
V147	xx11_6	xx11_6	contin	numeric
V148	xx11_7	xx11_7	contin	numeric
V149	xx11_8	xx11_8	contin	numeric
V150	xx11_9	xx11_9	contin	numeric
V151	xx11_10	xx11_10	contin	numeric
V152	xx11_11	xx11_11	contin	numeric
V153	xx11_12	xx11_12	contin	numeric
V154	xx11_13	xx11_13	contin	numeric
V155	xx11_14	xx11_14	contin	numeric
V156	xx11_99	xx11_99	contin	numeric
V157	co11	co11	discrete	numeric
V158	co12	co12	discrete	numeric
V159	so12_1	so12_1	discrete	numeric
V160	so12_2	so12_2	discrete	numeric
V161	so12_3	so12_3	discrete	numeric
V162	so12_99	so12_99	discrete	numeric
V163	hv01	hv01	discrete	numeric
V164	hs11	hs11	discrete	numeric
V165	hs01	hs01	discrete	numeric
V166	tm11	tm11	discrete	numeric
V167	tm12	tm12	discrete	numeric
V168	hi01	hi01	discrete	numeric
V169	hi03	hi03	discrete	numeric
V170	hi14	hi14	discrete	numeric
V171	ch11	ch11	discrete	numeric

V172	ch12	ch12	discrete	numeric
V173	ch13	ch13	discrete	numeric
V174	so14	s014	discrete	numeric
V175	so15	s015	discrete	numeric
V176	so16	s016	discrete	numeric
V177	so17	s017	discrete	numeric
V178	so18	s018	discrete	numeric
V179	hs12	hs12	discrete	numeric
V180	ar03_12	ar03_12	contin	numeric
V181	so11	so11	discrete	numeric
V182	ar03_13	ar03_13	contin	numeric
V183	ar03_99	ar03_99	contin	numeric
V184	ar03_100	ar03_100	contin	numeric
V185	hi12_99	hi12_99	contin	numeric
V186	hi11_99	hi11_99	discrete	numeric