South Africa - SAPRIN Individual Surveillance Episodes 2025 Datasets

Augustine Khumalo, Louis Sibiya, Dr Kobus Herbst, Tinofa Mutevedzi, Dr Molulaqhooa Linda Maoyi, Dr André Rose

Report generated on: October 7, 2026

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

Identification

SURVEY ID NUMBER
SAPRIN.SISED2025V1

TITLE
SAPRIN Individual Surveillance Episodes 2025 Datasets

COUNTRY

Name	Country code
South Africa	RSA

STUDY TYPE
Demographic Surveillance

SERIES INFORMATION
This dataset contains demographic surveillance data covering the period from 1 Jan 1993 to 31 December 2024.

ABSTRACT
The 'South African Population Research Infrastructure Network' (SAPRIN) is a national research infrastructure funded through the Department of Science, Technology and Innovation and hosted by the South African Medical Research Council. One of SAPRIN's initial goals has been to harmonise and share the longitudinal data from the four current Health and Demographic Surveillance System *(HDSS)* Nodes. These long-standing nodes are the MRC/Wits University Agincourt HDSS in Bushbuckridge District, Mpumalanga, established in 1993, with a current population of 119050 people; the University of Limpopo DIMAMO HDSS in the Capricorn District of Limpopo, established in 1996, with a current population of 103868; the Africa Health Research Institute (AHRI) HDSS in uMkhanyakude District, KwaZulu-Natal, established in 2000, with a current population of 159204; and the GRT-INSPIRED HDSS, Gauteng, established in 2020, with a current population of 68417.

For an individual to be eligible for inclusion in the surveillance, the individual must be a member of a household resident within the geographic boundaries of a SAPRIN node. For a household to be resident, it must have at least one household member who is resident within the surveillance area. Households and household membership are self-defined by the household informant interviewed by the fieldworker at their place of residence (or during a telephonic interview with the household informant). Household members so identified could be resident, that is sleep the majority of nights at this household's place of residence, or could be resident elsewhere (usually outside the surveillance area, but potentially within the surveillance area, in which case they will be a resident household member of the household resident at that place - In this dataset, individuals are members of a single household at a time, and in this example, the non-resident member of the household who is resident elsewhere in the surveillance area, will be reflected in the dataset as a resident member of the household this individual is co-resident with and not also as a non-resident member of this household). The resident status of household members can change: they can move out of the surveillance area to be resident elsewhere, but still be considered household members (so-called 'temporary migration'), such cases are reflected in the data as episodes of external residence; or temporary migrants can return to take up residence again with the household, initiating a new episode of residence internal to the surveillance area.

In addition to these periods of internal and external residence punctuated by in- and outmigration, surveillance episodes can be started by the birth of an individual, if the child is born to a resident mother, their birth starts a period of internal residency for the child; if the child is born to a mother who is a temporary migrant (externally resident) and the child is considered to be a member of the household, a period of external residency ensues for the child. Residency episodes (whether internal or external) are of course terminated by the death of the individual, if that happens whilst the individual is under surveillance.

All SAPRIN nodes conducted a baseline household census at their beginning and all individuals enrolled at this point start their surveillance episode with enumeration. However, nodes may extend their area of surveillance at certain points after the initial household census by doing another baseline census in these new areas, and all individuals enrolled then also start their surveillance episodes with enumeration. For integrity in the longitudinal surveillance of individuals, the identity of newly encountered individuals is checked against the database and merged with prior records if the individual is already in the database. In the case of newly incorporated areas into the surveillance area, it is entirely possible to find individuals that have previously out-migrated from the surveillance area to reside in this new area, in such cases an individual will have more than one surveillance episode that starts with enumeration, their enumeration in the original baseline census as well as their enumeration in the newly extended surveillance area.

This dataset represents a snapshot of the continually evolving data in the underlying longitudinal databases maintained by the SAPRIN nodes. In these databases the rightmost extend of the individual's surveillance episode is indicated by the data collection date of the last time the individual's membership of a household under surveillance has been confirmed. Each dataset has a right censor date (31 December 2024 for the current version of the dataset) and individual surveillance episodes are terminated at that point if the individual is still under surveillance beyond the cut-off date.

Each individual surveillance episode is associated with a physical location, for internal residency episodes it is the actual place of residence of the individual, for external residence episodes (periods of temporary migration) it is the place of residence of the individual's household. If an individual change their place of residency from one location within the surveillance area to another location still within the surveillance area, the episode at the original location is terminated with a location exit event, and a new episode starts with a location entry event at the destination location. It is also possible for the household the individual is a member of, to change their place of residency in the surveillance area, whilst the individual is externally resident (is a temporary migrant), in which case the individual's external resident episode will also be split with a location exit-entry pair of events.

At every household visit written consent is obtained from the household respondent for continued participation in the surveillance and such consent can be withdrawn. When this happens all household members' surveillance episodes are terminated with a refusal event. It is possible for households to again provide consent to participate in the surveillance after some time, in such cases surveillance events are restarted with a permission event. As mentioned previously, surveillance episodes are continually extended by the last data collection event if the individual remains under surveillance. In certain cases, individuals may be lost to follow-up and surveillance episodes where the date of last data collection is more than one year prior to the right censor data are terminated as lost to follow up at that last data collection date. Individuals with data collection dates within a year of the right censor date is considered still to be under surveillance up to this last data collection date.

Each surveillance episode contains the identifier of the household the individual is a member of during that episode. Under relatively rare circumstances it is possible for an individual to change household membership whilst still resident at the same location, or to change membership whilst externally resident, in these cases the surveillance episode will be split with a pair of membership end and membership start events. More commonly membership start and end events coincide with location exit and entry events or in- and out-migration events. Memberships also obviously start at birth or enumeration and end at death, refusal to participate or lost to follow-up.

In about half of the cases, individuals have a single episode from first enumeration, birth or in-migration, to their eventual death, out-migration or currently still under surveillance. In the remaining cases, individuals transition from internal residency to external residency via out-migration, or from one location to another via internal migration with a location exit and entry event, or some other rarer form of transition involving membership change, refusal or lost to follow-up. Usually these series of surveillance episodes are continuous in time, with no gaps between episodes, but gaps can form, e.g. when an individual out-migrates and end membership with the household and so is no longer under surveillance, only to return via in-migration at some future date and take up membership with same or different household.

Please take note that the migration event counts in the most recent years are inflated in two out of the three nodes by their operations not reconciling returning migrants or internal (within-site) migrants to their former records in the database. This has produced an artefactual increase in migration, but due to the design of SAPRIN's longitudinal data, this does not adversely affect population denominator or exposure measurements.

The SAPRIN Individual Surveillance Episodes 2025 Datasets consist of three types of the Demographic surveillance datasets:
1. SAPRIN Individual Surveillance Episodes 2025: Basic Dataset. This dataset contains only the internal and external residency episodes for an individual.
2. SAPRIN Individual Surveillance Episodes 2025: Age-Year Dataset. This dataset splits the basic surveillance episodes at calendar year end and at the date when the age in years (birth-day) of an individual changes.
3. SAPRIN Individual Surveillance Episodes 2025: Age-Year-Delivery Dataset. This dataset splits the basic surveillance episodes at calendar year end and at the date when the age in years (birth-day) of an individual changes. In the case of women who have given births, episodes are split at the time of delivery as well.

KIND OF DATA
Event history data

UNIT OF ANALYSIS
Exposure Episodes

Version

VERSION DESCRIPTION
v1: Dataset for public distribution.

VERSION DATE
2026-02-27

VERSION NOTES
v1: Dataset for public distribution.

Scope

NOTES
Each record in the 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. An out-migration will result in a new episode, because the location or residential status 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.

TOPICS

Topic
Fertility, Mortality, Migration

KEYWORDS

Keyword
Fertility, Mortality, Migration

Coverage

GEOGRAPHIC COVERAGE
The South African Population Research Infrastructure Network (SAPRIN) currently represents a network of four Health and Demographic Surveillance System *(HDSS)* nodes located in rural 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) and the Gauteng Research Triangle INSPIRED (GRT-I) HDSS in Gauteng Province, which has collected data since 2020. The nodal website is: https://grt.ac.za/inspired/. The Agincourt HDSS covers a surveillance area of approximately 420 square kilometres and is located in the Bushbuckridge District, Mpumalanga in the rural northeast of South Africa close to the Mozambique border. At baseline in 1992, 57 600 people were recorded in 8900 households in 20 villages; by 2006, the population had increased to about 70 000 people in 11 700 households. As of 1st July 2024, there were 118 381 people under surveillance of whom 35% were not resident within the surveillance area, with a total of about 2.9 m person years of observation. 30% of the population is under 15 years old. The population is almost exclusively Xitsonga speaking. The Agincourt HDSS has population density of over 200 persons per square kilometre. The Agincourt HDSS extends between latitudes 24° 50´ and 24° 56´S and longitudes 31°08´ and 31°´ 25´ E. The altitude is about 400-600m above sea level. DIMAMO is located in the Capricorn district, Limpopo Province approximately 40 kilometres from Polokwane, the capital city of Limpopo Province and 15-50 kilometres from the University of Limpopo. The site covers an area of approximately 400 square kilometres. The initial total population observed was about 8 000 but the field site was expanded in 2010. As of 1st July 2024, there were 103 238 people under surveillance, of whom 26% were not resident within the surveillance area, with about 1.1 m person years of observation. 28% of the population is under 15 years old. The population is predominantly Sepedi speaking. Most households have electricity. Some households have piped water either inside the house or in their yards, but most fetch water from taps situated at strategic points in the villages. Most households have a pit latrine in their yards. The area lies between latitudes and 23°65´ and 23°90´S and longitudes 29°65´ and 29°85´E. The HDSS is located on a high plateau area (approximately 1250 m above sea level) where communities typically consist of households clustered in villages, with access to local land

for small-scale food production.Africa Health Research Institute (AHRI) is situated in the south-east portion of the Umkhanyakude district of KwaZulu-Natal province near the town of Mtubatuba. It is bounded on the west by the Umfolozi-Hluhluwe nature reserve, on the south by the Umfolozi river, on the east by the N2 highway (except form portions where the Kwamsane township stradles the highway) and in the north by the Inyalazi river for portions of the boundary. The surveillance area is approximately 850 square kilometres. As of 1st July 2024, there were 160 302 people under surveillance of whom 32% were not resident within the surveillance area, with about 2.8 m person years of observation. 28 % of the population is under 15 years old. The population is almost exclusively isiZulu speaking. The surveillance area is typical of many rural areas of South Africa in that while predominantly rural, it contains an urban township and informal peri-urban settlements. The area is characterised by large variations in population densities (20-3000 people per square kilometre). The area lies between latitudes -28°24' and 28°20'N and longitudes 32°10' and 31°58'EGRT-INSPIRED operates across three primary surveillance areas situated within Gauteng Province: Atteridgeville, Melusi and Hillbrow. Atteridgeville is located west of Pretoria, bounded in the North by the WF Nkomo street. Attached to it, in the south-west, is an informal settlement called Jeffsville. Melusi is an informal settlement located north-west of Pretoria, near Kirkney and Booysens. It is bounded by the Daspoort ridge to the south, and the Van Der Hoff road (R514) to the north. Hillbrow, in the inner city of Johannesburg, is defined by high-rise residential buildings and is bounded by commissioner Street to the south, Clarendon PI street to the north, Klein Street to the west, and Nugget Street to the east, with a division between Hillbrow North and Hillbrow South (on Smit street) for operational purposes. Together, these areas form the GRT-INSPIRED surveillance node, encompassing a mixture of formal, informal, backyard, and high-density apartment dwellings across a combined high-population urban environment.

UNIVERSE

Households resident in dwellings within the study area will be eligible for inclusion in the household component of SAPRIN. All individuals identified by the household proxy informant as a member of the household will be enumerated. A resident household member is an individual that intends to sleep the majority of time at the dwelling occupied by the household over a four-month period. Households will include resident and non-resident members. An individual is a non-resident member if they have close ties to the household, but do not physically reside with the household most of the time. They can also be called temporary migrants and they are enumerated within the household list. Because household membership is not tied to physical residency, an individual may be a member of more than one household.

Producers and sponsors

PRIMARY INVESTIGATORS

Name	Affiliation
Augustine Khumalo	SAPRIN
Louis Sibiya	SAPRIN
Dr Kobus Herbst	SAPRIN
Tinofa Mutevedzi	SAPRIN
Dr Molulaqhooa Linda Maoyi	SAPRIN
Dr André Rose	SAPRIN

PRODUCERS

Name	Affiliation	Role
Dr Molulaqhooa Linda Maoyi	SAPRIN	Technical Assistance
Dr Kobus Hersbt	SAPRIN	Technical Assistance
Augustine Khumalo	SAPRIN	Technical Assistance
Louis Sibiya	SAPRIN	Technical Assistance

FUNDING AGENCY/SPONSOR

Name	Abbreviation	Role
Department of Science,Technology and Innovation	DSTI	Current Funder

OTHER IDENTIFICATIONS/ACKNOWLEDGMENTS

Name	Affiliation	Role
Agincourt Data Team	Agincourt	Providing Data
DIMAMO Data Team	DIMAMO	Providing Data
AHRI Data Team	AHRI	Providing Data
Prof Steve Tollman	Agincourt	
Dr Joseph Tlouyamma	DIMAMO	
Prof Collins Iwuji	AHRI	
Centre for High Performance Computing	Centre for High Performance Computing	Providing IT Infrastucture for Data Processing

Sampling

SAMPLING PROCEDURE

This dataset is not based on a sample but contains information from the complete demographic surveillance areas.

Data collection

DATES OF DATA COLLECTION

Start	End	Cycle
1993-01-01	2024-12-31	Agincourt
1996-01-01	2024-12-31	DIMAMO
2000-01-01	2024-12-31	AHRI
2020-01-01	2024-12-31	GRTI

TIME PERIODS

Start date	End date	Cycle
1993-01-01	2024-12-31	Agincourt
1996-01-01	2024-12-31	DIMAMO
2000-01-01	2024-12-31	AHRI
2020-01-01	2024-12-31	GRTI

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

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 SAPRINQA data quality assessment and indicator process. The data quality checking process is conducted using the SAPRIN QA Julia Code. The Julia Code 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 the SAPRIN Management hub is satisfied that they have high quality data. To produce this final standard dataset, the data is processed using Julia Code on the Centre for High Performance Computing cluster.

Data Appraisal

ESTIMATES OF SAMPLING ERROR
Not Applicable

DATA APPRAISAL
Not Applicable

Access policy

CONTACTS

Name	Affiliation	Email	URL
SAPRIN Data Management	SAPRIN	saprindata@mrc.ac.za	Link

ACCESS CONDITIONS
This data is made available for access under the following conditions:
1)The data and other materials provided by SAPRIN will not be redistributed or sold to other individuals, institutions, or organizations without the written agreement of SAPRIN.
2)The data will be used for statistical and scientific research purposes only. They will be used solely for reporting of aggregated information, and not for investigation of specific individuals or organisations. The Data User will neither use nor permit others to use the data in any way other than listed in the original application (Analysis Plan) for access to the dataset.
3)No attempt will be made to re-identify respondents, and no use will be made of the identity of any person or establishment discovered inadvertently. Any such discovery should immediately be reported to SAPRIN.
4)No attempt will be made to produce links among datasets provided by SAPRIN, or among data from SAPRIN and other datasets that could identify individuals or organizations.
5)The Data User will ensure that the data are kept in a secured environment and that only authorized users have access to the data.
6)Any books, articles, conference papers, theses, dissertations, reports, or other publications that employ data obtained from SAPRIN will cite the source of data in accordance with the Citation Requirement provided with each dataset.
7)An electronic copy of all reports and publications based on the requested data will be sent to SAPRIN.

8)The original collector of the data, SAPRIN, and relevant funding agencies bear no responsibility for use of the data or for interpretations or inferences based upon such uses.
9) Once the data set has served its indicated purpose it must be destroyed. If the dataset needs to be lodged for publication purposes, a reference (a digital object identifier will be maintained by SAPRIN for this purpose) to the original dataset on the SAPRIN data repository should be used. Derived or aggregated datasets produced from the original dataset do not fall within this provision and may be lodged as publication datasets. If the same dataset is needed for a different purpose, the dataset should be re-requested and the new purposes indicated.

CITATION REQUIREMENTS
Any use of this data must cite the digital object identifier (doi) associated with the appropriate dataset. Using the following form:

Maoyi, ML; Khumalo, AS; Sibiya, L; Mutevedzi, T; Rose, A (2025): SAPRIN Individual Surveillance Episodes 2025: Basic Dataset. South African Population Research Infrastructure Network.https://doi.org/10.23667/saprin.sisebd2025

Maoyi, ML; Khumalo, AS; Sibiya, L; Mutevedzi, T; Rose, A (2025) : SAPRIN Individual Surveillance Episodes 2025: Age-Year Dataset. South African Population Research Infrastructure Network. https://doi.org/10.23667/saprin.siseayd2025

Maoyi, ML; Khumalo, AS; Sibiya, L; Mutevedzi, T; Rose, A (2025) : SAPRIN Individual Surveillance Episodes 2025: Age-Year-Delivery Dataset. South African Population Research Infrastructure Network. https://doi.org/10.23667/saprin.siseaydd2025

ACCESS AUTHORITY

Name	Affiliation	Email	URL
SAPRIN Data Management	SAPRIN	saprindata@mrc.ac.za	Link

Disclaimer and copyrights

DISCLAIMER
The user of the data acknowledges that the original collector of the data and the relevant funding agencies bear no responsibility for the data's use or interpretation or inferences based upon it.

COPYRIGHT
This dataset documentation is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License. The dataset is shared in terms of the data-use agreement accepted at the time of data download.

Metadata production

DDI DOCUMENT ID
DDI.SAPRIN.SISED2025V1

PRODUCERS

Name	Abbreviation	Affiliation	Role
Molulaqhooa Linda Maoyi	MLM	SAPRIN	Documentation of Study and Review of the metadata
Augustine Khumalo	AK	SAPRIN	Documentation of Study and Review of the metadata
Lous Sibiya	LS	SAPRIN	Documentation of Study and Review of the metadata

DATE OF METADATA PRODUCTION
2026-02-27

DDI DOCUMENT VERSION
Version 1 (February 2026)

Data Dictionary

Data file	Cases	Variables
SISEDB2025V1	1519530	35
SISEAYD2025V1	15076425	41
SISEAYDD2025V1	15196634	42