# South Africa - SAPRIN Partnership, Pregnancy and HIV Observations 2025 Dataset

**Dr Andre Rose, Tinofa Mutevedzi, Dr Molulaqhooa Linda Maoyi, Augustine Khumalo**

Report generated on: October 27, 2025

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

# Identification

SURVEY ID NUMBER
SAPRIN.SPPHOD2025V1

TITLE
SAPRIN Partnership, Pregnancy and HIV Observations 2025 Dataset

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 2015 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 three 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 104,679 people; the University of Limpopo DIMAMO HDSS in the Capricorn District of Limpopo, established in 1996, with a current population of 101,376; and the Africa Health Research Institute (AHRI) HDSS in uMkhanyakude District, KwaZulu-Natal, established in 2000, with a current population of 156,751.
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. This involves complex tracking of internal and external residency episodes, in- and out-migration events, births, deaths, and changes in household membership and consent status. All of these events structure the individual's surveillance history within the underlying longitudinal HDSS databases. Each surveillance episode is continually extended by the last data collection event if the individual remains under surveillance. The underlying longitudinal databases have a right censor date (31 December 2024 for the current version), at which point individual surveillance episodes are terminated if the individual is still under surveillance.
The SAPRIN Partnership, Pregnancy and HIV Observations 2025 Dataset
The SAPRIN Partnership, Pregnancy and HIV Observation 2025 Dataset is a longitudinal episode dataset derived from the SAPRIN HDSS databases. It presents the continuous history of surveillance for female individuals whose episode records fall within the observation period of 2015-2024.
Each record in this dataset represents a specific surveillance episode for a female individual, a time period during which all recorded characteristics remained constant. This structure allows researchers to measure time-at-risk and exposure duration for various outcomes.
Key variables included for each episode are:
· NodeId
· Deidentified Individual ID
· Date of Birth
· Observation Date (The end date of the episode, or date of the status being recorded)
· Age
· Partnership Status
· Pregnancy Status
· HIV Ever Tested
· HIV Result
This harmonized longitudinal resource is powerful for studying the dynamics and temporal associations of demographics, partnership status, pregnancy, and HIV status among women across the three SAPRIN population sites.

KIND OF DATA
Event history data

UNIT OF ANALYSIS
Status Observations

## Version

VERSION DESCRIPTION
v1: Dataset for public distribution.

VERSION DATE
2025-10-27

VERSION NOTES
v1: Dataset for public distribution.

## Scope

NOTES
The SAPRIN Partnership, Pregnancy and HIV Observation 2025 Dataset is a longitudinal episode dataset containing the continuous surveillance history for female individuals between the ages of 18 and 24 across the Agincourt, DIMAMO, and AHRI HDSS nodes between 2015 and 2024.
Unit of Analysis
The fundamental unit of analysis is the observation. Each record represents a single observation of a female aged 18-24 years within the surveillance period (2015-2024), capturing her demographic and health characteristics at that specific point in time.

TOPICS

| Topic |
| --- |
| HIV, Partnership, Pregnancy |

KEYWORDS

| Keyword |
| --- |
| HIV, Partnership, Pregnancy |

## Coverage

GEOGRAPHIC COVERAGE
The South African Population Research Infrastructure Network (SAPRIN) currently represents a network of three 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) and 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.

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 2023, there were 110 896 people under surveillance of whom 24% were not resident within the surveillance area, with a total of about 2.8 m person years of observation. 29% 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 2023, there were 99 087 people under surveillance, of whom 24% were not resident within the surveillance area, with about 961,000 person years of observation. 29% 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 2023, there were 154 815 people under surveillance of whom 25% were not resident within the surveillance area, with about 2.6 m person years of observation. 29 % 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 characterized 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'E

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 |
|---|---|
| Dr Andre Rose | SAPRIN |
| Tinofa Mutevedzi | SAPRIN |
| Dr Molulaqhooa Linda Maoyi | SAPRIN |
| Augustine Khumalo | SAPRIN |

PRODUCERS

| Name | Affiliation | Role |
|---|---|---|
| Dr Molulaqhooa Linda Maoyi | SAPRIN | Technical Assistance |
| Augustine Khumalo | 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 represents a sample of individuals drawn from the full demographic surveillance populations of the three SAPRIN Health and Demographic Surveillance System *(HDSS)* nodes - Agincourt, DIMAMO, and AHRI.
Selection Criteria

The SAPRIN Partnership, Pregnancy and HIV Observation 2025 Dataset was created by selecting a subsample of individuals based on the following demographic and temporal criteria:
Sex: Only female individuals were included.
Age at Observation: Individuals aged 18 to 24 years (inclusive) during the period of interest were selected.
Timeframe: Inclusion was limited to those observed between 2015 and 2024 within the HDSS datasets.

All available variables including nodeid, individualid, age, partnershipstatus, pregnancystatus, hivevertested, and hivresult are provided for each individual meeting these criteria.

The resulting dataset therefore represents a targeted sample of young adult women (18-24 years) observed across the three SAPRIN HDSS nodes between 2015 and 2024.

## 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 |

FREQUENCY OF DATA COLLECTION
Three rounds per year.

TIME PERIODS

| Start date | End date | Cycle |
|---|---|---|
| 2015-01-01 | 2024-12-31 | Agincourt |
| 2015-01-01 | 2024-12-31 | DIMAMO |
| 2015-01-01 | 2024-12-31 | AHRI |

DATA COLLECTION MODE
CAPI and CATI

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; Mutevedzi, T; Rose, A (2025): SAPRIN Partnership, Pregnancy and HIV Observations 2025 Dataset. South African Population Research Infrastructure Network.

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



## Metadata production

DDI DOCUMENT ID

DDI.SAPRIN.SPPHOD2025V1

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 |

DATE OF METADATA PRODUCTION

2025-10-27

DDI DOCUMENT VERSION

Version 1 (October 2025)

## Data Dictionary

| Data file | Cases | Variables |
|---|---|---|
| **SPPHOD2025V1** | 0 | 10 |

**Data file: SPPHOD2025V1**

| | |
|---|---|
| Cases: | 0 |
| Variables: | 10 |

**Variables**

| ID | Name | Label | Question |
|---|---|---|---|
| V1 | NodeId | Study Node | |
| V2 | IndividualId | Individual Identifier | |
| V3 | Sex | Sex | |
| V4 | DateOfBirth | Date of Birth | |
| V5 | ObservationDate | Observation Date | |
| V6 | Age | Age at Observation | |
| V7 | PartnershipStatus | Partnership Status | |
| V8 | PregnancyStatus | Pregnancy Status | |
| V9 | HIVEverTested | HIV Ever Tested | |
| V10 | HIVResult | HIV Test Result | |

Total: 10

## NODEID: Study Node

**Data file: SPPHOD2025V1**

### Overview

Valid: 0 Invalid: 0
Type: Discrete Decimal: 0 Width: 9 Range: 1 - 3 Format: Numeric

### Questions and instructions

CATEGORIES

| Value | Category |
|---|---|
| 1 | Agincourt |
| 2 | DIMAMO |
| 3 | AHRI |

## INDIVIDUALID: Individual Identifier

**Data file: SPPHOD2025V1**

### Overview

Valid: 0 Invalid: 0
Type: Continuous Decimal: 0 Width: 9 Range: 1 - 7518 Format: Numeric

## SEX: Sex

**Data file: SPPHOD2025V1**

### Overview

Valid: 0 Invalid: 0
Type: Discrete Decimal: 0 Width: 12 Range: 0 - 3 Format: Numeric

### Questions and instructions

CATEGORIES

| Value | Category |
|---|---|
| 0 | None/Missing |
| 1 | Male |
| 2 | Female |
| 3 | Unknown |

## DATEOFBIRTH: Date of Birth

**Data file: SPPHOD2025V1**

### Overview

Valid: 0
Type: Discrete Width: 11 Range: - Format: character

## OBSERVATIONDATE: Observation Date

**Data file: SPPHOD2025V1**

### Overview

Valid: 0
Type: Discrete Width: 11 Range: - Format: character

## AGE: Age at Observation

**Data file: SPPHOD2025V1**

### Overview

Valid: 0 Invalid: 0
Type: Discrete Decimal: 0 Width: 8 Range: 18 - 24 Format: Numeric

## PARTNERSHIPSTATUS: Partnership Status

**Data file: SPPHOD2025V1**

### Overview

Valid: 0 Invalid: 0
Type: Discrete Decimal: 0 Width: 19 Range: 0 - 4 Format: Numeric

### Questions and instructions

CATEGORIES

| Value | Category |
|---|---|
| 0 | Missing/Refused |
| 1 | No partnership |
| 2 | Marital partnership |
| 3 | Regular partnership |
| 4 | Casual partnership |

## PREGNANCYSTATUS: Pregnancy Status

**Data file: SPPHOD2025V1**

### Overview

Valid: 0 Invalid: 0
Type: Discrete Decimal: 0 Width: 10 Range: 0 - 4 Format: Numeric

### Questions and instructions

CATEGORIES

| Value | Category |
|---|---|
| 0 | Undefined |
| 1 | Yes |
| 2 | No |
| 3 | Don't Know |
| 4 | Not asked |

## HIVEVERTESTED: HIV Ever Tested

**Data file: SPPHOD2025V1**

### Overview

Valid: 0 Invalid: 0
Type: Discrete Decimal: 0 Width: 9 Range: 0 - 2 Format: Numeric

### Questions and instructions

CATEGORIES

| Value | Category |
|---|---|
| 0 | No |
| 1 | Yes |
| 2 | Unknown |

## HIVRESULT: HIV Test Result

**Data file: SPPHOD2025V1**

### Overview

Valid: 0 Invalid: 0
Type: Discrete Decimal: 0 Width: 13 Range: 0 - 3 Format: Numeric

### Questions and instructions

CATEGORIES

| Value | Category |
|---|---|
| 0 | Negative |
| 1 | Positive |
| 2 | Indeterminate |
| 3 | Not tested |