South Africa

Dr. Kobus Herbst, AHRI, SAPRIN
Dr. Guy Harling, AHRI
Dr. Mark Siedner, AHRI
Prof Willem Hanekom, AHRI

AHRI.Education Cleaned Dataset

Study Documentation

February 18, 2026

Metadata Production

Metadata Producer(s)	Africa Health Research Institute (AHRI)
Identification	DDI.AHRI.EducationCleaned

Table of Contents

AHRI.Education Cleaned Dataset

Overview

Identification	AHRI.EducationCleaned
Version	v2.0.0

Abstract

Africa Health Research Institute (AHRI) is a multidisciplinary, independent research institute. AHRI's goal is to become a source of fundamental discoveries into the susceptibility, transmission and cure of HIV, TB and related diseases. We also seek ways to improve diagnosis, prevention and treatment. To achieve this, we bring together leading researchers from different fields, use cutting edge science to improve people's health, and help to train the next generation of African scientists.

AHRI has conducted individual and household demographic surveillance since 2000. Demographic surveillance data are processed for longitudinal analyses by organizing the demographic surveillance data into individual and household residency episodes at a geographical physical location, and individual membership episodes within a household. Start events of residency and memberships include enumeration, birth, in-migration and relocating into a household from within the study population. Exit events of residency and memberships include death (by cause), out-migration and relocating to another location in the study population. AHRI annually produces a suite of Individual Demographic surveillance datasets which contains detailed records of all individuals under surveillance since January 2000. The datasets detail individual residency and membership patterns, household mortality (including cause of death), fertility, migration, HIV status, health care utilization, and household socio-economic profile.

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, marital status, social-economic status and HIV will likewise result in a new episode on the date of the change.

The following are the versions of the Demographic surveillance datasets:

CoresidencyEpisodes
EducationCleaned
EducationStatuses
HIVStatusEpisodes
HouseholdAssets
HouseholdMap
HouseholdObservations
HouseholdSocioEconomicStatus
HouseholdStatusEpisodes
IndividualHSEEpisodes
IndividualMap
LabourEpisodes
LabourStatuses
LocationMap
MaritalStatuses
PartnerEpisodes
SurveillanceEpisodesBasic
SurveillanceEpisodesYTAge
SurveillanceEpisodes YrAge_Coresidency
SurveillanceEpisodesYrAge_HIV
SurveillanceEpisodes VrAge_HSE
SurveillanceEpisodes YrAge_Labour
SurveillanceEpisodesYrAge_Partners
SurveillanceEpisodesYrAgeDel

SurveillanceEpisodes VrAgeDel_HSE
SurveillanceEpisodesYrAgeDeL_HSE_Labour
SurveillanceEpisodesYrAgeDeL_HSE_Labour_Partner_Parents
SurveillanceEpisodes VrAgeDel_HSE_Labour_Partner_Parents_HIV
SurveillanceEpisodesYrAgeDeL_HSE_Labour_Partners

Kind of Data	Event History Data
Unit of Analysis	Episodes of exposure

Scope & Coverage

Keywords	Fertility, Mortality, Migration, Social-economic status, HIV, ART
Topics	Education Cleaned
Time Period(s)	2000-2024
Countries	South Africa

Geographic Coverage
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 strandles the highway) and in the north by the provincial road R22. The surveillance area is approximately 845km2 in size.

Universe
Households resident in dwellings within the study area are eligible for inclusion in the household component of AHRI Demographic Surveillance. All individuals identified by the household proxy informant as members of the households are enumerated. A resident household member is an individual that intends to sleep majority of time at the dwelling occupied by the household over a four-month period. Households 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 at the dwelling most of the time. Non-resident members 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 & Sponsors

Primary Investigator(s)	Dr. Kobus Herbst, AHRI, SAPRIN Dr. Guy Harling, AHRI Dr. Mark Siedner, AHRI Prof Willem Hanekom, AHRI
Other Producer(s)	Africa Health Research Institute (AHRI)
Funding Agency/ies	Wellcome Trust (WT) , Core funding SAPRIN (SAPRIN)
Other Acknowledgment(s)	AHRI Research Data Management Team , Research Data Management , AHRI AHRI Data Collection Team , Data Collection , AHRI AHRI Community Engagement Team , Community Engagement , AHRI

Sampling

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

Data Collection

Data Collection Dates	start 2000-01-01 end 2024-12-31

Data Processing & Appraisal

Data Editing

Data collected in this study was centrally stored on AHRI's secure database servers. Data was collected through interviews at the call centre on central server using Survey Solutions, and in the field on paper and encrypted tablet computers using REDCap, and then securely uploaded to a central server. Database rules and constraints were enforced on the master database to ensure that only valid data was stored. Personally identifiable data was restricted to separate database tables accessible only to data management staff .

Accessibility

Access Conditions

Access to the data requires accurate completion of the online data access application form accessible on the AHRI Data repository(<https://data.ahri.org/>). Data users are required to abide by the data use conditions stipulated on the application for access to the data. Failure to do so may result in their data access privileges being revoked by the Data Custodian. In order to recognise the effort and intellectual contributions of AHRI investigators in producing and curating the data, users of AHRI data must acknowledge the source of the data and abide by the terms and conditions under which the data is accessed and must cite the dataset in publication using the citation provided as part of this documentation. All analytical datasets published on the AHRI Data Repository are assigned digital object identifier (DOIs) and the DOIs can be found on the Data Repository under Study Description tab - Access policy. AHRI data users are required to always cite the dataset using the relevant DOI.

https://github.com/SAPRIN-MRC/SaprinEpisodes.jl.git

Citation Requirements

Herbst, K., Harling, G., Siedner, M., & Hanekom, W. (2026). AHRI.Education Cleaned Dataset [Data set]. Africa Health Research Institute.

DOI:https://doi.org/10.23664/AHRI.EDUCATIONCLEANED

Files Description

Dataset contains 1 file(s)

EducationCleaned	
# Cases	3046012
# Variable(s)	6

Variables List

Dataset contains 6 variable(s)

File EducationCleaned

#	Name	Label	Type	Format	Valid	Invalid	Question
1	Individu ..	Individual ID	continuous	numeric-9.0	3046012	0	-
2	Year	Year in which education level was attained	continuous	numeric-9.0	3046012	0	-
3	DoB	Date of birth of individual	discrete	character-11	3046004	-	-
4	Age	Age attained during this year	continuous	numeric-9.0	3046012	0	-
5	Educatio ..	School level attained in this year prior to cleaning	discrete	numeric-18.0	1737652	1308360	-
6	CleanedE ..	School level attained in this year after to cleaning	discrete	numeric-18.0	2195404	850608	-

Variables Description

Dataset contains 6 variable(s)

File : EducationCleaned

IndividualId: Individual ID

Information	[Type= continuous] [Format=numeric] [Range= 1-263153] [Missing=*]
Statistics [NW/ W]	[Valid=3046012 /-] [Invalid=0 /-] [Mean=119070.693 /-] [StdDev=77552.346 /-]

Year: Year in which education level was attained

Information	[Type= continuous] [Format=numeric] [Range= 2000-2024] [Missing=*]
Statistics [NW/ W]	[Valid=3046012 /-] [Invalid=0 /-] [Mean=2013.473 /-] [StdDev=7.308 /-]

DoB: Date of birth of individual

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=3046004 /-]

Age: Age attained during this year

Information	[Type= continuous] [Format=numeric] [Range= 0-1500] [Missing=*]
Statistics [NW/ W]	[Valid=3046012 /-] [Invalid=0 /-] [Mean=25.58 /-] [StdDev=18.777 /-]

EducationStatus: School level attained in this year prior to cleaning

Information	[Type= discrete] [Format=numeric] [Range= -1-99] [Missing=*]
Statistics [NW/ W]	[Valid=1737652 /-] [Invalid=1308360 /-]

Value	Label	Cases	Percentage
-1	Unknown	86515	5.0%
0	Grade 0	31894	1.8%
1	Grade 1	59802	3.4%
2	Grade 2	66632	3.8%
3	Grade 3	71886	4.1%
4	Grade 4	80926	4.7%
5	Grade 5	69139	4.0%
6	Grade 6	73859	4.3%
7	Grade 7	96034	5.5%
8	Grade 8	92848	5.3%
9	Grade 9	102759	5.9%
10	Grade 10	147361	8.5%
11	Grade 11	173666	10.0%
12	Grade 12	484021	27.9%
13	Special School	0	
21	Higher Certificate	0	
22	National Diploma	0	
23	Degree	0	
24	Honours	0	
25	Masters	0	
26	Doctors	0	
31	ABET 1	0	
32	ABET 2	0	
33	ABET 3	0	
34	ABET 4	0	
41	FET 1	0	

File : EducationCleaned

EducationStatus: School level attained in this year prior to cleaning

Value	Label	Cases	Percentage
42	FET 2	0	
43	FET 3	0	
50	LT 1 year	0	
51	Creche	0	
96	Refused	0	
97	NAD	0	
98	No education	100310	5.8%
99	Not attending	0	
Sysmiss		1308360	

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.

CleanedEducation: School level attained in this year after to cleaning

Information	[Type= discrete] [Format=numeric] [Range= -1-99] [Missing=*]
Statistics [NW/ W]	[Valid=2195404 /-] [Invalid=850608 /-]

Value	Label	Cases	Percentage
-1	Unknown	38468	1.8%
0	Grade 0	32154	1.5%
1	Grade 1	67377	3.1%
2	Grade 2	87930	4.0%
3	Grade 3	89473	4.1%
4	Grade 4	114764	5.2%
5	Grade 5	85390	3.9%
6	Grade 6	92041	4.2%
7	Grade 7	122007	5.6%
8	Grade 8	119479	5.4%
9	Grade 9	113892	5.2%
10	Grade 10	180246	8.2%
11	Grade 11	182734	8.3%
12	Grade 12	696221	31.7%
13	Special School	0	
21	Higher Certificate	0	
22	National Diploma	0	
23	Degree	0	
24	Honours	0	
25	Masters	0	
26	Doctors	0	
31	ABET 1	0	
32	ABET 2	0	
33	ABET 3	0	
34	ABET 4	0	
41	FET 1	0	
42	FET 2	0	
43	FET 3	0	
50	LT 1 year	0	

File : EducationCleaned

CleanedEducation: School level attained in this year after to cleaning

Value	Label	Cases	Percentage
51	Creche	0	
96	Refused	0	
97	NAD	0	
98	No education	173228	7.9%
99	Not attending	0	
Sysmiss		850608	

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.