South Africa - Africa Health Research Institute INDEPTH Core Dataset 2000-2015 (Residents only) - Release 2017

Kobus Herbst - Africa Health Research Institute (ZA031), Frank Tanser - Africa Health Research Institute (ZA031), Deenan Pillay - Africa Health Research Institute (ZA031)

Report generated on: June 30, 2017

Overview

Identification

ID NUMBER
INDEPTH.ZA031.CMD2015.v1

Version

VERSION DESCRIPTION
1.0

PRODUCTION DATE
2017-06-29

Overview

ABSTRACT

The health and demography of the South African population has been undergoing substantial changes as a result of the rapidly progressing HIV epidemic. Researchers at the University of KwaZulu-Natal and the South African Medical Research Council established The Africa Health Research Studies in 1997 funded by a core grant from The Wellcome Trust, UK. Given the urgent need for high quality longitudinal data with which to monitor these changes, and with which to evaluate interventions to mitigate impact, a demographic surveillance system (DSS) was established in a rural South African population facing a rapid and severe HIV epidemic. The DSS, referred to as the Africa Health Research Institute Demographic Information System (ACDIS), started in 2000.

ACDIS was established to 'describe the demographic, social and health impact of the HIV epidemic in a population going through the health transition' and to monitor the impact of intervention strategies on the epidemic. South Africa's political and economic history has resulted in highly mobile urban and rural populations, coupled with complex, fluid households. In order to successfully monitor the epidemic, it was necessary to collect longitudinal demographic data (e.g. mortality, fertility, migration) on the population and to mirror this complex social reality within the design of the demographic information system. To this end, three primary subjects are observed longitudinally in ACDIS: physical structures (e.g. homesteads, clinics and schools), households and individuals. The information about these subjects, and all related information, is stored in a single MSSQL Server database, in a truly longitudinal way—i.e. not as a series of cross-sections.

The surveillance area is located near the market town of Mtubatuba in the Umkanyakude district of KwaZulu-Natal. The area is 438 square kilometers in size and includes a population of approximately 85 000 people who are members of approximately 11 000 households. The population is almost exclusively Zulu-speaking. The 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/km2). In the rural areas, homesteads are scattered rather than grouped. Most households are multi-generational and range with an average size of 7.9 (SD:4.7) members. Despite being a predominantly rural area, the principle source of income for most households is waged employment and state pensions rather than agriculture. In 2006, approximately 77% of households in the surveillance area had access to piped water and toilet facilities.

To fulfil the eligibility criteria for the ACDIS cohort, individuals must be a member of a household within the surveillance area but not necessarily resident within it. Crucially, this means that ACDIS collects information on resident and non-resident members of households and makes a distinction between membership (self-defined on the basis of links to other household members) and residency (residing at a physical structure within the surveillance area at a particular point in time). Individuals can be members of more than one household at any point in time (e.g. polygamously married men whose wives maintain separate households). As of June 2006, there were 85 855 people under surveillance of whom 33% were not resident within the surveillance area. Obtaining information on non-resident members is vital for a number of reasons. Most importantly, understanding patterns of HIV transmission within rural areas requires knowledge about patterns of circulation and about sexual contacts between residents and their non-resident partners. To be consistent with similar datasets from other INDEPTH Member centres, this data set contains data from resident members only.

During data collection, households are visited by fieldworkers and information supplied by a single key informant. All births, deaths and migrations of household members are recorded. If household members have moved internally within the surveillance area, such moves are reconciled and the internal migrant retains the original identfier associated with him/her.

KIND OF DATA
Event history data

UNITS OF ANALYSIS
Individual

Scope

NOTES
This study represents only a portion of the total data associated with the complete AHRI Population Intervention Platform as described in the study abstract.

It specifically only includes the events defining the resident exposure of individuals under surveillance as well as the delivery events of resident women. Each type of event contains minimal attributes describing the event:

Attributes common to each event:

Event Type,

Event Date

Observation date

Migration:

Origin & Destination

Death:

Cause

Delivery:

Live born and Still born counts

Parity

TOPICS

Topic	Vocabulary	URI
Demography [N01.224]	MeSH	http://www.ncbi.nlm.nih.gov/mesh
Age Distribution [N01.224.033]	MeSH	http://www.ncbi.nlm.nih.gov/mesh
Emigration and Immigration [N01.224.625.350]	MeSH	http://www.ncbi.nlm.nih.gov/mesh
Residential Mobility [N01.224.791.700]	MeSH	http://www.ncbi.nlm.nih.gov/mesh
Sex Distribution [N01.224.803]	MeSH	http://www.ncbi.nlm.nih.gov/mesh
Vital Statistics [N01.224.935]	MeSH	http://www.ncbi.nlm.nih.gov/mesh
Life Expectancy [N01.224.935.464]	MeSH	http://www.ncbi.nlm.nih.gov/mesh
Mortality [N01.224.935.698]	MeSH	http://www.ncbi.nlm.nih.gov/mesh
Cause of Death [N01.224.935.698.100]	MeSH	http://www.ncbi.nlm.nih.gov/mesh
Birth Rate [N01.224.935.849.500]	MeSH	http://www.ncbi.nlm.nih.gov/mesh
Rural Population [N01.600.725]	MeSH	http://www.ncbi.nlm.nih.gov/mesh
Maternal Age [N06.850.490.250.550]	MeSH	http://www.ncbi.nlm.nih.gov/mesh

Topic	Vocabulary	URI
Parity [N06.850.490.812.600]	MeSH	http://www.ncbi.nlm.nih.gov/mesh
Survival Analysis [N06.850.520.830.998]	MeSH	http://www.ncbi.nlm.nih.gov/mesh

Coverage

GEOGRAPHIC COVERAGE

Demographic surveillance area 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 Inyalazi river for portions of the boundary.

The area is 438 square kilometers.

UNIVERSE

Resident household members of households resident within the demographic surveillance area. Inmigrants are defined by intention to become resident, but actual residence episodes of less than 180 days are censored. Outmigrants are defined by intention to become resident elsewhere, but actual periods of non-residence less than 180 days are censored. Children born to resident women are considered resident by default, irrespective of actual place of birth. The dataset contains the events of all individuals ever resident during the study period (1 Jan 2000 to 31 Dec 2015)

Producers and Sponsors

PRIMARY INVESTIGATOR(S)

Name	Affiliation
Kobus Herbst	Africa Health Research Institute (ZA031)
Frank Tanser	Africa Health Research Institute (ZA031)
Deenan Pillay	Africa Health Research Institute (ZA031)

OTHER PRODUCER(S)

Name	Affiliation	Role
Tinofa Mutevedzi	Africa Health Research Insittute (ZA031)	Data Collection

FUNDING

Name	Abbreviation	Role
Wellcome Trust	WT	Current Funder
Wellcome Trust	WT	Prior Funder

OTHER ACKNOWLEDGEMENTS

Name	Affiliation	Role
Dickman Gareta	Africa Health Research Institute (ZA031)	Database Scientist
Sweetness Dube	Africa Health Research Institute (ZA031)	Data Documentation Archivist

Metadata Production

METADATA PRODUCED BY

Name	Abbreviation	Affiliation	Role
iSHARE2 Technical Team	iS2TT	INDEPTH Network	Documentation of the study
INDEPTH Network	int.indepth	INDEPTH Network	Agency

Name	Abbreviation	Affiliation	Role
AJ Herbst	AJH	ZA031	DDI author
SH Dube	SHD	ZA031	Data Documentation Archivist

DATE OF METADATA PRODUCTION
2017-06-29

DDI DOCUMENT VERSION
Version 1 (June 2017)

DDI DOCUMENT ID
DDI.INDEPTH.ZA031.CMD2015.v1

Sampling

Sampling Procedure

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

Reponse units (households) by year:

Year	Households
2000	11856
2001	12321
2002	12981
2003	12165
2004	11841
2005	11312
2006	12065
2007	12165
2008	11790
2009	12145
2010	12485
2011	12455
2012	12087
2013	11988
2014	11778
2015	11938

In 2006 the number of response units increased due to the addition of a new village into the demographic surveillance area.

Deviations from Sample Design

None

Response Rate

Household response rates are as follows (assuming that if a household has not responded for 2 years following the last recorded visit to that household, that the household is lost to follow-up and no longer part of the response rate denominator)

Year	Response Rate
2000	94%
2001	93%
2002	96%
2003	91%
2004	88%
2005	84%
2006	88%
2007	89%
2008	87%
2009	88%
2010	89%
2011	89%
2012	89%
2013	90%
2014	89%
2015	91%

Weighting

Not applicable

Questionnaires

Overview

List of questionnaires

Bounded structure registration (BSR) or update (BSU) form
• used to register characteristics of the BS
• Updates characteristics of the BS
• Information as at previous round is preprinted

Household registration (HHR) or update (HHU) form
• used to register characteristics of the HH
• Used to update information about the composition of the household
• Information preprinted of composition and all registered households as at previous.

Household Membership Registration (HMR) or update (HMU)
• used to link individuals to households.
• Used to update information about the household memberships and member status observations
• Information preprinted of member status observations as at previous.

Individual registration form (IDR)
• Used to uniquely identify each individual
• Mainly to ensure members with multiple household memberships are appropriately
captured

Migration notification form (MGN)
• Used to record change in the BS of residency of individuals or households
• Migrants are tracked and updated in the database

Pregnancy history form (PGH) & pregnancy outcome notification form (PON)
• Records details of pregnancies and their outcomes
• Only if woman is a new member
• Only if woman has never completed WHL or WGH

Death notification form (DTN)
• Records all deaths that have recently occurred
• includes information about time, place, circumstances and possible cause of death

Data Collection

Data Collection Dates

Start	End	Cycle
2000-01-01	2015-12-31	Release coverage

Time Periods

Start	End	Cycle
2000-02-01		Round 1
2000-08-01		Round 2
2001-02-01		Round 3
2001-06-25		Round 4
2002-01-07		Round 5
2002-05-06		Round 6
2002-09-02		Round 7
2003-01-10		Round 8
2003-06-26		Round 9
2003-12-01		Round 10
2004-07-01		Round 11
2005-01-02		Round 12
2005-07-01		Round 13
2006-01-02		Round 14
2006-07-18		Round 15
2007-01-06		Round 16
2007-07-02		Round 17
2008-01-14		Round 18
2008-07-01		Round 19
2009-01-01		Round 20
2009-07-13		Round 21
2010-01-03		Round 22
2010-06-16		Round 23
2011-01-02		Round 24
2011-07-02		Round 25
2012-01-12		Round 26
2012-05-25		Round 27
2012-09-03		Round 28
2013-01-02		Round 29
2013-08-20		Round 31
2014-01-04		Round 32
2014-05-01		Round 33
2014-08-21		Round 34
2015-01-02		Round 35
2015-05-02		Round 36
2015-08-22		Round 37

Data Collection Mode

Proxy Respondent [proxy]

DATA COLLECTION NOTES

Enumerators were trained immediately prior to the baseline data collection and then refresher training was conducted for one week between each surveillance round. New fieldworkers received a standardised 6 week training course prior to appointment as data collectors. Data entry staff received fieldwork training in addition to training in the use of the data entry programs.

Data Collectors

Name	Abbreviation	Affiliation
The Africa Health Research Institute	ZA031	UKZN

SUPERVISION

Fieldworkers operated in teams of between 8 and 12 fieldworkers supervised each supervised by a Fieldwork supervisor. Supervisors conduct supervised visits and quality control visits and review fieldworkers data collection.

Data Processing

Data Editing

On data entry data consistency and plausibility were checked by 455 data validation rules at database level. If data validaton failure was due to a data collection error, the questionnaire was referred back to the field for revisit and correction. If the error was due to data inconsistencies that could not be directly traced to a data collection error, the record was referred to the data quality team under the supervision of the senior database scientist. This could request further field level investigation by a team of trackers or could correct the inconsistency directly at database level.

No imputations were done on the resulting micro data set, except for:

a. If an out-migration (OMG) event is followed by a homestead entry event (ENT) and the gap between OMG event and ENT event is greater than 180 days, the ENT event was changed to an in-migration event (IMG).
b. If an out-migration (OMG) event is followed by a homestead entry event (ENT) and the gap between OMG event and ENT event is less than 180 days, the OMG event was changed to an homestead exit event (EXT) and the ENT event date changed to the day following the original OMG event.
c. If a homestead exit event (EXT) is followed by an in-migration event (IMG) and the gap between the EXT event and the IMG event is greater than 180 days, the EXT event was changed to an out-migration event (OMG).
d. If a homestead exit event (EXT) is followed by an in-migration event (IMG) and the gap between the EXT event and the IMG event is less than 180 days, the IMG event was changed to an homestead entry event (ENT) with a date equal to the day following the EXT event.
e. If the last recorded event for an individual is homestead exit (EXT) and this event is more than 180 days prior to the end of the surveillance period, then the EXT event is changed to an out-migration event (OMG)

In the case of the village that was added (enumerated) in 2006, some individuals may have outmigrated from the original surveillance area and setlled in the the new village prior to the first enumeration. Where the records of such individuals have been linked, and indivdiual can legitmately have and outmigration event (OMG) forllowed by and enumeration event (ENU). In a few cases a homestead exit event (EXT) was followed by an enumeration event in these cases. In these instances the EXT events were changed to an out-migration event (OMG).

Other Processing

All homesteads in the Hlabisa sub-district were geocoded and entered into a geographic information system (GIS) prior to the start of surveillance. The demographic surveillance area was selected on the basis of this information to include an area with clear geographic boundaries and an estimated population size suitable for the envisaged research agenda. Since then the GIS database has been updated based on notification of new homesteads from the fieldwork and periodic reviews of satellite and aerial photography.

Mapping teams used differentially coorrected global positioning system (GPS) units (accuracy <2m) to geocode homesteads.

How document control was conducted to ensure all census forms were completed?

Before each round, a SQL script generated a list of questionnaires to be printed for each household resident in the surveillance area. Each questionnaire is given a unique integer key which is printed as a barcode on the questionnaire. A series of web-based reports called 'Unified Reports' are then used to track and control the status of each questionnaire from document production, data collection, data entry and document archiving. A strict chain of custody is enforced for all questionnaire movements.

A data entry is performed by a team of 6 data capturers with one supervisor using in-house developed software (Delphi and .NET C#). Double-entry is not routinely used except in the case of verbal autopsy questionnaires.

Data is stored in a MS SQL database, with transaction logging, daily backups and twice weekly off-site backups. Constraints and validation rules placed on the database help in checking data quality during data entry.

All data entry done by each data capturer in the first five days of each round is 100% rechecked by the supervisor. If during those 5 days the data capturer's work is consistently error-free, only 20% of their work will be subjected to rechecking by a supervisor. If any error is picked up in the 20% rechecking, then their work gets subjected to 100% recheck for another 5 consecutive days.

Field QC Procedures

- Supervised visits - this exercise is carried by the fieldworker and the supervisor jointly. The two select a sample of bounded structures which they will visit together. During a Supervised visit, the supervisor listens and observes as the fieldworker conducts the interviews without interrupting. The supervisor uses a checklist to write observations and comments for feedback and further training of a particular fieldworker immediately after departure from a BS,. The supervised visit checklist is submitted to the QC section and is used for performance analysis, as well as for identification of training needs.

- Quality Control visits - these are repeat data collection visits conducted by a fieldwork supervisor soon after the fieldworker completes routine data collection at a homestead. This is done mainly to ensure accuracy and reliability of the information collected by fieldworkers. Quality control visits are selected randomly by the computer at a 5% sample of the total number of homesteads to be visited per each round. The original copy and the supervisor's copy are then compared by the quality controllers to identify discrepancies between the two. If discrepancies are found, the two copies are rejected back to the field for reconciliation between the two. The records are also kept at the quality control section for analyses towards the end of the round and this also contributes to performance management of individual employees QC at the office before data entry.

After data collection and before data entry, the office-based QC section checks questionnaires for completeness, consistency and accuracy. If a questionnaire failed to meet the quality standard requirements, the QC clerk send back the questionnaires to the field worker's supervisor.

Specify how the data was extracted (including which software program was used) to produce the core micro data set. How was inconsistent records dealt with during this process?

Following data collection and data entry completion at the end of a surveillance round, a snapshot of the operational database is created as an analytical database. each such snapshot is uniquely identified and analytical datasets must reference the analytical database thay originated from. Analytical datasets are never produced directly from the operational database, as this database is continually in flux as data is updated through the data collection and entry processes.

An sql script produces a normalised episode table each time an analytical database is created. This episode table contains an exposure record for each exposure episode for an individual, from initial enumeration, birth or in-migration, up to eventual death or out-migration. The episode table contains the start event and date of the exposure as well as the end event and date of the end of exposure. Individuals that out-migrate and later in-migrate are reconciled as far as possible using individual identifiers (national identity number, names, sex and date of birth) under a single individual identity. All internal movements (migrations) are reconciled and residencies at different homesteads within the surveillance area are reflected as separate episodes in the episode table.

In the case of deaths, the next of kin are visited by a verbal autopsy nurse and a derivation of the INDEPTH standard verbal autopsy questionnaire is used to document the death. The verbal autopsy questionnaires are interpreted by the INTERVA-4 program to derive cause of death information.

To produce this micro-data set, the episode table is processed using Pentaho Kettle ETL program to produce this standard event-history format dataset.

Data Appraisal

Estimates of Sampling Error

Not applicable

File Description

Variable List

ZA031.CMD2015.v1

Content	Event History Micro Data Set
Cases	559222
Variable(s)	14
Structure	Type: Keys: ()
Version	CMD2015.v1
Producer	Africa Health Research Institute
Missing Data	

Variables

ID	Name	Label	Type	Format	Question
V1	RecNr	RecNr	discrete	numeric	
V2	CountryId	CountryId	discrete	numeric	
V3	CentreId	CentreId	discrete	character	
V4	IndividualId	IndividualId	discrete	numeric	
V5	Sex	Sex	discrete	numeric	
V6	DoB	DoB	discrete	character	
V7	EventCount	EventCount	discrete	numeric	
V8	EventNr	EventNr	discrete	numeric	
V9	EventCode	EventCode	discrete	character	
V10	EventDate	EventDate	discrete	character	
V11	ObservationDate	ObservationDate	discrete	character	
V12	LocationId	LocationId	discrete	numeric	
V13	MotherId	MotherId	discrete	numeric	
V14	DeliveryId	DeliveryId	discrete	numeric	

RecNr (RecNr)

File: ZA031.CMD2015.v1

Overview

Type: Discrete
Format: numeric
Decimals: 0
Range: 1-488221

Valid cases: 559222
Invalid: 0

Description

A sequential number uniquely identifying each record in the data file

CountryId (CountryId)

File: ZA031.CMD2015.v1

Overview

Type: Discrete
Format: numeric
Decimals: 0
Range: 710-710

Valid cases: 559222
Invalid: 0

Description

ISO 3166-1 numeric code of the country in which the surveillance site is situated

CentreId (CentreId)

File: ZA031.CMD2015.v1

Overview

Type: Discrete
Format: character
Width: 5

Valid cases: 559222
Invalid: 0

Description

An identifier issued by INDEPTH to each member centre of the format CCCSS, where CCC is a sequential centre identifier and SS is a sequential identifier of the site within the centre in the case of multiple site centres

IndividualId (IndividualId)

File: ZA031.CMD2015.v1

Overview

Type: Discrete
Format: numeric
Decimals: 0
Range: 1-135061

Valid cases: 559222
Invalid: 0

Description

A number uniquely identifying all the records belonging to a specific individual in the data file. This number is not be the same as the identifier used by a contributing centre to identify the individual.

Sex (Sex)

File: ZA031.CMD2015.v1

Overview

Sex (Sex)

File: ZA031.CMD2015.v1

Type: Discrete
Format: numeric
Decimals: 0
Range: 0-2

Valid cases: 559222
Invalid: 0

Description

Sex of the individual.

DoB (DoB)

File: ZA031.CMD2015.v1

Overview

Type: Discrete
Format: character

Valid cases: 559222
Minimum: NaN
Maximum: NaN

Description

The date of birth of the individual. Format: YYYY/MM/DD

EventCount (EventCount)

File: ZA031.CMD2015.v1

Overview

Type: Discrete
Format: numeric
Decimals: 0
Range: 2-22

Valid cases: 559222
Invalid: 0

Description

The total number of events associated with this individual in this data set

EventNr (EventNr)

File: ZA031.CMD2015.v1

Overview

Type: Discrete
Format: numeric
Decimals: 0
Range: 1-22

Valid cases: 559222
Invalid: 0

Description

A number increasing from 1 to EventCount for each event record in order of event occurrence

EventCode (EventCode)

File: ZA031.CMD2015.v1

Overview

Type: Discrete
Format: character
Width: 3

Valid cases: 559222
Invalid: 0

Description

A code identifying the type of event that has occurred.

EventDate (EventDate)
File: ZA031.CMD2015.v1

Overview

Type: Discrete
Format: character

Valid cases: 559222
Minimum: NaN
Maximum: NaN

Description

The date on which the event occurred. Format: YYYY/MM/DD

ObservationDate (ObservationDate)
File: ZA031.CMD2015.v1

Overview

Type: Discrete
Format: character

Valid cases: 558575
Minimum: NaN
Maximum: NaN

Description

Date on which the event was observed (recorded), also known as surveillance visit date. Format: YYYY/MM/DD

Dates less than 1800/01/01 are mssing values

LocationId (LocationId)
File: ZA031.CMD2015.v1

Overview

Type: Discrete
Format: numeric
Decimals: 0
Range: 1-14774

Valid cases: 549729
Invalid: 9493

Description

Unique identifier associated with a residential unit within the site and is the location where the individual was or became resident when the event occurred. This identifier is not be the same as the identifier used internally by the contributing centre.

MotherId (MotherId)
File: ZA031.CMD2015.v1

Overview

Type: Discrete
Format: numeric
Decimals: 0
Range: 1-135068

Valid cases: 25727
Invalid: 533495

Description

The IndividualId of the mother. Only provided for BTH events.

DeliveryId (DeliveryId)
File: ZA031.CMD2015.v1

Overview

DeliveryId (DeliveryId)

File: ZA031.CMD2015.v1

Type: Discrete	Valid cases: 25727
Format: numeric	Invalid: 533495
Decimals: 0	
Range: 1-17359	

Description

The RecNr of the delivery event associated with this birth