

Africa Centre Demographic Information System (ACDIS)

Analytical Dataset Documentation

Demography Dataset

Change History

<u>Who</u>	<u>When</u>	<u>What</u>
Kobus Herbst, Colin Newell	12 October 2009	Original
Colin Newell	18 th January 2010	Updated for 2009 data. Added Balance Sheet example analysis
Colin Newell	20 th January 2010	Added birth and death rate analysis example
Colin Newell	26 th May 2010	Added <i>EarliestHIVNegative</i>
Colin Newell	7 th July 2010	Updated using 30 June 10 dataset data

Dataset name	ACDISDemography
Description	<p>This dataset contains a detailed record of all individuals under surveillance between 1 Jan 2000 and June 2010. It details their residency pattern, mortality (incl. cause of death), fertility, migration and HIV status.</p> <p>Each person's time under surveillance is split into numerous "exposure episodes" and there is one row per exposure episode in the dataset. Episodes are specific to a) Individual, b) semester, c) Age Group and d) Residency Type. Episodes span at most a six-month 'semester' i.e. the ten years 2000-2009 are split into 20 semesters, Jan-Jun and Jul-Dec.</p> <p>If an individual changes their Age Group during a semester (e.g. say they have their 15th birthday on 1st February) then two episodes are created – one spanning the part of the semester they were in the 10-14 age group, i.e. covering 1st Jan up until 31st Jan and another from when they joined the 15-19 age group i.e. from 1st February to 30th June. Similarly, if the individual's Residency was to change at, say, 1st April, then that would also cause separate episodes to be created. Hence one individual can have over 30 episodes.</p>
Rows/Units of analysis	Episodes of exposure
Unique identifier(s)	Reclid
Approx. number of rows	Nearly 2.26 million as at June 2010

Introduction

Individual exposure episodes are accumulated from:

- a. Start of surveillance, 1 Jan 2000, if the individual was resident or a non-resident household member at the start of surveillance, flag *PeriodStart* = 1
- b. From date of birth, if the date of birth is after 1 Jan 2000 and the individual was resident from birth, flag *AgedIn*=1
- c. From start of first household membership episode, if the individual was non-resident at the start of the household membership, flag *MemberStart* = 1
- d. From the start of the first residential episode of the individual, if the individual was not resident from birth or a household member prior to in migration, flag *InMigrEx* = 1

The flags *PeriodStart* and *MemberStart* can only appear on an individual's first episode. The flag *InMigrEx* can appear on subsequent exposure episodes, but only if the individual migrated out of the surveillance area and later in-migrated again. The flag *AgedIn* will be set (=1) at the start of each episode where the individual was under surveillance prior to this episode and the episode start coincides with the individual aging into a particular age group. It will also be set if the start of surveillance coincides with the birth of the individual. In other words counting births under surveillance is equal to summing *AgedIn* where *Episode*=1

Exposure episodes continue to accumulate if the individual out migrates but retains membership of a resident household.

Each exposure episode is marked as:

- a. Resident (*EpisodeType* = 1) if the individual was resident during that episode
- b. Non-resident (*EpisodeType* = 2) if the individual was a member of at least one household during that period, and was a non-resident, and there was a previous Resident episode.
- c. Non-resident (*EpisodeType* = 3) if the individual was a member of at least one household prior and this was prior to them first becoming resident. i.e. there is no prior episode with *EpisodeType* = 1

For *EpisodeType* = 1, the *BSIntID* of the bounded structure where the individual was resident is recorded. If *EpisodeType* = 1 and the *BSIntID* is missing, it implies that there was an internal migration, but that the new residence of the individual was not recorded.

Typically, mortality, fertility rates etc are calculated only using episodes of Type 1 and 2. The basic rationale is that we know very little about individuals until they become resident for the first time. Mortality is typically only calculated for Residents (Type 1) as their exposure to the determinants of mortality is likely to differ for non-residents.

Exposure episodes end:

- a. If the individual is still under surveillance at the end of the surveillance period, flag *PeriodEnd*=1

- b. If the individual dies before or at the end of the surveillance period, flag *Died=1*
- c. If the individual externally out migrated (either individually or as part of a household) and did not retain membership of any household, flag *OutMigrEx=1*
- d. If the visit at which the status of the individual was last recorded is a surveillance visit prior to the end of the surveillance period, *Lost=1*. If the individual's residency ends on an internal migration and this is the last date prior to *PeriodEnd* on which the individual was observed then *Lost=1* and *OutMigr=1*. In other words these are individuals who internally migrated and were not subsequently tracked.
- e. If the last episode of the individual is non-resident (*EpisodeType* 2 or 3) and a membership end (Event Type HDS or HEM) is recorded prior to *PeriodEnd*, flag *MemberEnd=1*

The start (*ObservationStart*) and end (*ObservationEnd*) dates of each surveillance episode are inclusive. The days of exposure (*ExpDays*) in each episode = $(\text{ObservationEnd} - \text{ObservationStart}) + 1$

Demography Dataset Variable List

Variable Name	Description	Format/Codes/Special values	Notes/comments
<i>ReclD</i>	Unique record Id,	Integer 1 – 2.2million+	Increases sequentially per individual and episode number
<i>IIntID</i>	Individual internal identifier	Integer 11-153605	
<i>Sex</i>	Sex of individual	FEM Female MAL Male MIS Missing	Only 5 individuals (30 episodes) are MIS, mainly early neonatal deaths
<i>Episodes</i>	Total number of exposure episodes for this individual	Integer 1 to34	Constant within each <i>IIntID</i>
<i>Episode</i>	Episode number for this individual	Integer 1 to 34	Sequential 1 to <i>Episodes</i> within each <i>IIntID</i> Episodes start/end at semester, age-group and residency (episode) type boundaries To get last episode use: “WHERE <i>Episode</i> = <i>Episodes</i> ”
<i>EpisodeType</i>	Type of episode	1 Residency episode 2 Non-residency episode following a residency episode. 3 Non-residency episode prior to any residency episode.	Episodes are split on <i>EpisodeType</i> Typically demographic rates exclude episodes of Type 3, sometimes also Type 2
<i>BSIntID</i>	Internal ID of the residential Bounded Structure	Integer 11-approx 16375 Null	Null for all non-resident episodes (i.e. where <i>EpisodeType</i> is 2 or 3) and also for resident episodes following an internal migration where the destination is unknown.
<i>ExpYear</i>	Year in which this episode falls i.e. exposure occurred	Integer 2000 to 2009	

Demography Dataset Variable List

Variable Name	Description	Format/Codes/Special values	Notes/comments
<i>ExpSem</i>	Semester of <i>ExpYear</i> in which this episode falls i.e. in which exposure occurred	1 In first half of year (i.e. Jan-Jun) 2 In second half of year (i.e. Jul – Dec)	
<i>AgeGrp</i>	Age Group into which the individual fell during all of this episode	Integer 1-23 1 0-27 days (i.e.neonates) 2 28 days up to 1 year 3 1-2 years 4 2-3 years 5 3-4 years 6 4-5 years 7 5-9 years 8 to 22: 10-14 up to 80-84 years, in 5yr intervals 23 85+ years	Because episodes are split on age group as well as semester the <i>AgeGrp</i> is an attribute of the whole episode, not its start or end.
<i>ObservationStart</i>	Exposure episode start date	Date between 1 Jan 2000 and Dec 2009	
<i>ObservationEnd</i>	Exposure episode end date	Date between 1 Jan 2000 and Dec 2009	The difference between <i>ObservationStart</i> and <i>ObservationEnd</i> is always between 1 and 183 days
<i>ExpDays</i>	Number of days of exposure in episode	Integer 1 to 184	Number of days between <i>ObservationStart</i> and <i>ObservationEnd</i> , inclusive of both start and end date. Hence it will be 1 if <i>ObservationStart</i> equals <i>ObservationEnd</i> date.
<i>PeriodStart</i>	Was under surveillance at start of surveillance period	0 No 1 Yes	95% are 0=No Yes can only appear on <i>Episode 1</i> except in the case of Indlovu village exposure, which was added to the Surveillance Area in 2006. Indlovu exposure has thus been left-censored to 1 Oct 2006.

Demography Dataset Variable List

Variable Name	Description	Format/Codes/Special values	Notes/comments
<i>AgedIn</i>	Aged into episode	0 No 1 Yes	86% are 0=No Yes for <i>Episode</i> 1 indicates the individual was born into observation. Yes for any other Episode indicates the individual aged into the episode.
<i>MemberStart</i>	The episode starts at the start of Household membership	0 No 1 Yes	99.8% are 0=No This can only be Yes when <i>Episode</i> = 1 and <i>EpisodeType</i> = 3
<i>InMigr</i>	Is the episode start due to a change in residence?	0 No 1 Yes	Is set to 1=Yes for both internal and external in and outmigrations.
<i>InMigrEx</i>	Is the episode start due to immigration from outside the DSA	0 No 1 Yes	97% are 0=No
<i>Died</i>	Does this episode end with the death of the individual?	0 Alive 1 Died	99.4% are alive
<i>MemberEnd</i>	Episode ends due to end of household membership	0 No 1 Yes	99.3% are 0=No Can only be Yes when <i>EpisodeType</i> is 2 or 3
<i>OutMigr</i>	Migrated out of current residence	0 No 1 Yes	94% are 0=No
<i>OutMigrEx</i>	Migrated out of DSA	0 No 1 Yes	97% are 0=No Can only be Yes when <i>EpisodeType</i> = 1
<i>Lost</i>	Lost to follow-up	0 No 1 Yes	99.2% are 0=No Is 1=Yes if the last observation of the individual is a Visit prior to the end of the observation period. Can only be set 1 on last Episode
<i>PeriodEnd</i>	Under surveillance at end of period	0 No 1 Yes	96% are No

Demography Dataset Variable List

Variable Name	Description	Format/Codes/Special values	Notes/comments
<i>PositiveExp</i>	Number of days observed as HIV +ve	Integer 0 to 183. No Nulls	
<i>NegativeExp</i>	Number of days observed as HIV -ve	Integer 0 to 183. No Nulls	
<i>UnknownExp</i>	Number of days observed with unknown HIV status	Integer 0 to 183. No Nulls	
<i>HIVPositive</i>	Is individual known to be HIV +ve during this episode	0 No 1 Yes	98% are 0= No
<i>HIVNegative</i>	Is individual known to be HIV -ve during this episode	0 No 1 Yes	82% are No If <i>HIVPositive</i> and <i>HIVNegative</i> are both 1=Yes, then the individual is known to have seroconverted during this episode.
<i>EarliestHIVPositive</i>	Earliest date on which the individual was known to be HIV positive	Date October 2002 to June 2010 Null if not known positive	This is the same for all episodes for an individual.
<i>EarliestHIVNegative</i>	Date of earliest HIV -ve test	Date Feb 2001 to June 2010	This is the same for all episodes for an individual. (Added 26may10)
<i>LatestHIVNegative</i>	Latest date on which the individual was known to be HIV negative	Date Feb 2001 to June 2010	This is the same for all episodes for an individual.
<i>Area</i>	Area of residence	0 Rural 1 Peri-urban 2 Urban 3 Outside surveillance area Null	Is Null if the individual is resident but not linked to any BoundedStructure i.e. unlinked internal migrations)
<i>C_Unknown</i>	Died of unknown cause during (i.e. at end of) this episode	0 No 1 Yes	

Demography Dataset Variable List

Variable Name	Description	Format/Codes/Special values	Notes/comments
<i>C_CMPN</i>	Died of communicable, maternal, perinatal or nutritional cause during (i.e. at end of) this episode	0 No 1 Yes	
<i>C_AIDS_TB</i>	Died of AIDS or TB during (i.e. at end of) this episode	0 No 1 Yes	
<i>C_NonComm</i>	Died of a non-communicable disease during (i.e. at end of) this episode	0 No 1 Yes	
<i>C_Injuries</i>	Died of injuries during (i.e. at end of) this episode	0 No 1 Yes	
<i>LBCnt</i>	Live birth count – The number of live births to this woman during this episode	0 No birth 1 Single birth 2 Twins 3 Triplets	Males, children etc get 0

Example Analyses

This section gives some examples of how to go about analysing the dataset. Initially they are simple examples, with code shown in both SQL and Stata, It is hoped this section will be added to over time as analysts submit more examples. You are encouraged to do so.

1. Balance sheet

The numbers under surveillance (Row a) in table below) are augmented by b) births, c) migrations into the DSA and d) by the addition of new non-resident members i.e. individuals who do are not resident in the DSA but are reported by resident households as being members of their household. Similarly the numbers will be reduced by e) deaths, f) outmigrations, g) non-resident members ending their membership and by h) losses to follow-up.

Hence this exercise computes the number under surveillance at the start of the year, adds on the births, immigrations etc, that occurred during the year and subtracts the deaths, outmigrations etc, and thereby ends up with i) the number under surveillance at the very end of the year.

The annual numbers, calculated from the 30 June 2010 database, are:

All Individuals (incl non-residents)	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Under surveillance at start of year	-	90,428	90,219	90,069	90,079	90,048	90,315	92,183	91,332	91,028
Start of surveillance	88,331	-	-	-	-	-	1,467	-	-	-
Born	2,207	2,060	2,055	1,972	1,990	2,050	1,942	1,972	1,881	1,575
Started Membership	1,318	546	402	281	261	310	360	250	321	295
Immigrated during year	3,057	3,049	4,414	3,261	2,619	2,531	2,533	2,400	2,082	2,121
Died during year	1,269	1,476	1,523	1,597	1,484	1,404	1,272	1,332	1,204	1,147
Membership end	1,712	2,054	2,310	1,554	1,272	1,318	1,372	1,430	1,199	917
Outmigrated during year	1,119	1,588	2,080	1,335	1,081	761	607	952	566	692
Lost to follow-up during year	385	746	1,108	1,018	1,064	1,140	1,167	1,759	1,619	2,005
Under surveillance at end of year	90,428	90,219	90,069	90,079	90,048	90,315	92,183	91,332	91,028	90,258
Discrepancy between years (e.g. End 200 and start 2001)	0	0	0	0	0	0	0	0	0	#REF!
Start + ins - outs	90,428	90,219	90,069	90,079	90,048	90,316	92,199	91,332	91,028	90,258
Discrepancy within columns	0	0	0	0	0	1	16	0	0	0

a. Under surveillance at start of year

Essentially this is a count of episodes that begin 1st January (recall that episodes are always split by semester). However, we have to be very careful to exclude from this count the births, migrations etc that occurred on 1st January itself as they are accounted for in the other numbers. We thus want the number under surveillance precisely as the clock strikes midnight of New Year's Eve.

```
SQL: SELECT
      ExpYear, Count(*)
FROM ACDISDemography
WHERE DAY(ObservationStart) = 1      -- episode starts 1st of month
      AND MONTH(ObservationStart) = 1 -- episode starts January
      AND NOT (AgedIn = 1 and Episode = 1) -- exclude those born on 1st Jan (as they are
                                          accounted for in the 'born' row)
      AND MemberStart = 0            -- exclude those becoming non-resident Members on 1st Jan
      AND NOT (InMigrEx = 1 and Episode = 1) -- exclude those migrating in for first time on 1st Jan
GROUP BY ExpYear
ORDER BY ExpYear
```

```
Stata: tab ExpYear if day(ObservationStart) == 1 & month(ObservationStart) == 1 &
      (AgedIn == 1 & Episode == 1) == 0 & MemberStart == 0 &
      (InMigrEx ==1 & Episode == 1) == 0
```

b. *Born during year*

Episodes beginning with birth always have *Episode* = 1 and *AgedIn* = 1, so we are just counting those.

```
SQL: SELECT ExpYear, COUNT(*)
FROM ACDISDemography
WHERE Episode = 1
      AND AgedIn = 1
GROUP BY ExpYear
```

```
Stata: tab ExpYear if Episode == 1 & AgedIn == 1
```

c. Started Membership

Variable *MemberStart* is set to 1 only for those episodes which start at the start of Household Membership. These will always have *Episode = 1*

```
SQL: SELECT ExpYear, COUNT(*)
      FROM ACDISDemography
      WHERE MemberStart = 1
      GROUP BY ExpYear
```

```
Stata: tab ExpYear if MemberStart == 1
```

d. Immigrated during year

Note that here we only want those where the immigration to the DSA is the individual's first episode (i.e. *Episode = 1*). We do not want those who start membership and then later migrate in, nor those who migrated out, and are now returning. The variable used to indicate that an episode starts with immigration from outside the DSA is *InMigrEx = 1*

```
SQL: SELECT ExpYear, COUNT(*)
      FROM ACDISDemography
      WHERE Episode = 1
            AND InMigrEx = 1
      GROUP BY ExpYear
```

```
Stata: tab ExpYear if Episode == 1 & InMigrEx == 1
```

e. Died during the year

Those episodes that end with death have variable *Died = 1*. These will always be on the last episode, (*Episode = Episodes*) but it is sufficient just to use the *Died* variable alone.

```
SQL: SELECT ExpYear, COUNT(*)
      FROM ACDISDemography
      WHERE Died = 1
      GROUP BY ExpYear
```

```
Stata: tab ExpYear if Died == 1
```

f. *Membership ended*

This is very similar indeed to the deaths row, except the crucial variable is *MemberEnd*. It is set = 1, only ever on an individual's last episode, when a non-resident person ceases to be under surveillance because their Household Membership is ended.

```
SQL: SELECT ExpYear, COUNT(*)
      FROM ACDISDemography
      WHERE MemberEnd = 1
      GROUP BY ExpYear
```

```
Stata: tab ExpYear if MemberEnd==1
```

g. *Outmigrated*

Here the critical variable is *OutMigrEx* – It is set to 1 if the individual (who must be a Resident i.e. *EpisodeType* = 1)) migrated out of the DSA. However, we also need to ensure this is on the individual's last episode as we wish to exclude outmigrations were the individual returned, or when they became non-resident Members.

```
SQL: SELECT ExpYear, COUNT(*)
      FROM ACDISDemography
      WHERE Episode = Episodes -- last episode for an individual
      AND OutMigrEx = 1
      GROUP BY ExpYear
```

Stata: tab ExpYear if Episode == Episodes & OutMigrEx == 1

h. Lost

Again the query is virtually identical, except the variable of interest is *Lost*, which is set to 1 (only ever on last *Episode*) when a person has been lost to follow-up.

SQL: `SELECT ExpYear, COUNT(*)
FROM ACDISDemography
WHERE Lost = 1
GROUP BY ExpYear`

Stata: tab ExpYear if Lost==1

i. Numbers under surveillance at year-end.

Here, as before, we want the numbers as at midnight so we need to be careful to exclude from the total those who died or outmigrated etc on 31st December itself.

SQL: `SELECT
ExpYear, Count(*)
FROM ACDISDemography
WHERE DAY(ObservationEnd) = 31 -- episode ends 31st of month
AND MONTH(ObservationEnd) = 12 -- episode ends December
AND Died = 0 -- exclude those dying on 31st Dec
AND MemberEnd = 0 -- exclude those ending non-resident
Membership on 31st Dec
AND NOT(OutMigrEx = 1 AND Episode = Episodes) -- exclude those migrating out on 31st Dec
AND Lost = 0 -- exclude those lost on 31 dec
GROUP BY ExpYear
ORDER BY ExpYear`

Stata: tab ExpYear if day(ObservationEnd) == 31 & month(ObservationEnd) == 12 &

```

Died == 0 &
MemberEnd == 0 &
(OutMigrEx ==1 & Episode == Episodes) == 0 &
Lost == 0

```

2. Crude Birth and Death Rates

While the CBR and CDR are conventionally calculated using the mid-year population as denominator, this is just because it is normally an easily-available proxy for the actual number of person-years of exposure to the risk of being born or dying during the year. In ACDIS, however, because it is a truly longitudinal dataset, exposure times are readily available and hence it is easier and better to calculate the actual exposure. Broadly, one sums the variable *ExpDays* and divides by 365.25 to turn days into years, then multiplies by 1,000 as conventionally the CBR and CDR are expressed per 1,000.

To calculate the numbers of deaths in an interval one just needs to sum the variable *Died*, because each dead individual has *Died* = 1 on their last episode. Similarly for births, though they are identified by having *Episode* = 1 and *AgedIn* = 1.

The annual CBR and CDR figures are given in the tables below. Note that, per convention, the births are calculated including episodes of both Type 1 and 2 i.e. we include both episodes of residency, and also episodes of non-residency that occur after a first episode of residency, but we omit episodes of non-residency that occur before any episode of residency. Put another way you must have resided in the area before you are included. Conversely, deaths are calculated, again as per convention, only for resident episodes. All episodes of non-residency (*EpisodeType* = 1 or 2) are excluded.

CBR	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
a)Births	1,686	1,745	1,733	1,663	1,767	1,751	1,699	1,733	1,654	1,382
b)Exposuredays	23,695,422	25,570,365	27,262,062	28,095,496	28,823,361	29,211,686	29,797,100	30,530,609	30,677,756	30,753,222
c)CrudeBirthRate	26.0	24.9	23.2	21.6	22.4	21.9	20.8	20.7	19.7	16.4

To calculate the births:

```

SQL: SELECT
      ExpYear, count(*)
FROM ACDISDemography
WHERE EpisodeType IN (1, 2)      -- episodes of residency plus non-res episodes after residency
      and Episode = 1 AND AgedIn = 1
GROUP BY ExpYear

```

```
ORDER BY ExpYear
```

```
Stata: tab ExpYear if (EpisodeType == 1 | EpisodeType == 2) & Episode == 1 & AgedIn == 1
```

And to calculate the exposure time

```
SQL: SELECT
      ExpYear,
      SUM(ExpDays) as ExpDays
FROM ACDISDemography
WHERE EpisodeType IN (1, 2)      -- episodes of residency plus non-res episodes after residency
GROUP BY ExpYear
ORDER BY ExpYear
```

```
Stata: tabstat ExpDays if (EpisodeType == 1 | EpisodeType == 2),
      by(ExpYear) stat(sum) format(%12.0gc)
```

CDR	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
a)deaths	889	1,097	1,134.0	1179	1045	1040	921	956	847	829
b)Exposuredays	23592767	24,849,605	25,095,072	24708711	24458977	24162880	24220698	24768852	24614449	24423556
c)CrudeBirthRate	13.8	16.1	16.5	17.4	15.6	15.7	13.9	14.1	12.6	12.4

For the count of deaths we have:

```
SQL: SELECT
      ExpYear, COUNT(*)
FROM ACDISDemography
WHERE EpisodeType = 1 -- episodes of residency only
      AND Died = 1
GROUP BY ExpYear
ORDER BY ExpYear
```

```
Stata: tab ExpYear if EpisodeType == 1 & Died == 1
```

While for the exposure days we have

```
SQL: SELECT
      ExpYear, COUNT(*)
FROM ACDISDemography
WHERE EpisodeType = 1 -- episodes of residency only
      AND Died = 1
GROUP BY ExpYear
ORDER BY ExpYear
```

```
Stata: tabstat ExpDays if EpisodeType == 1, by(ExpYear) stat(sum) format(%12.0gc)
```