INDEPTH Network

Better Health Information for Better Health Policy

Public Launching of …

INDEPTH Data Repository & INDEPTHStats

On 1st July 2013

A. INDEPTH Data Repository

What is the INDEPTH Data Repository?

The INDEPTH Data Repository is an online archive of various fully documented, high-quality datasets from INDEPTH member HDSS centres. Its goal is to enable INDEPTH member HDSSs and associated researchers to contribute and share HDSS datasets with the scientific community in support of the Network's mission. Every dataset is documented, using an internationally accepted metadata standard developed by the Data Documentation Initiative (DDI), enabling data users to quickly identify and obtain the data they require. Through the use of digital object identifiers (doi) the documentation promotes the citing of data sets by data users and facilitates the recognition of the efforts by the INDEPTH Network to make this valuable resource of population and health data for low- and middle-income countries (LMICs) available.

Why INDEPTH Data Repository?

"Science is based on building on, reusing and openly criticising the published body of scientific knowledge. For science to effectively function, and for society to reap the full benefits from scientific endeavours, it is crucial that science data be made open[1]." In this statement the authors of the "Panton Principles" elucidated the fundamental basis for the expectation from a growing body of scientists, research funders, and indeed the general public, that the data underlying published research or generated by research projects should be freely accessible.

Sharing research data can add value to the data at little cost by, for example, using the data to answer questions not anticipated by the original research, or facilitating meta-analyses by pooling similar data from many sources. Individuals who participate in and taxpayers who directly or

[1] Panton Principles, Principles for open data in science. Murray-Rust, Peter; Neylon, Cameron; Pollock, Rufus; Wilbanks, John; (19 Feb 2010). Retrieved 23 Apr 2013 from http://pantonprinciples.org/

indirectly support research are justified in expecting that the maximum benefit will be obtained from the research.

In response to a recent joint statement[2] by funders on the sharing of public health research data, the INDEPTH Executive Director, stressed in a reply[3], the importance of ensuring that the means and capacity to share and actively participate in the analysis of those data are in the hands of those who generate the data and not only in those who want to analyse it.

The INDEPTH Data Repository will be the first data repository that specialises in longitudinal individual level data from LMICs and presents an opportunity to exert a powerful and sustained influence on the availability of well documented and high quality longitudinal individual level exposure and cause specific mortality data from LMICs where such data has traditionally been difficult to obtain. The lack of research data management skills and capacity is generally acknowledged to be a major factor in why such data is not more generally available. As far back as 2006 the INDEPTH Network has taken important steps to address this through the establishment of the ISHARE project with core support funding from Sida/Research Cooperation Unit, Hewlett Foundation, Rockefeller Foundation and Wellcome Trust. ISHARE has recently been expanded into ISHARE2 as part of a Strategic Award funding from the Wellcome Trust, to:

- Enhance the research data management capacity of INDEPTH member centres and enable them to develop, document, extract, harmonise and quality-assure analytical datasets from their operational databases.
- Lead a major expansion in INDEPTH data sharing by establishing and maintaining this INDEPTH Data Repository.
- Strengthen and maintain the skills, procedures and infrastructure necessary to assure quality and responsibly share longitudinal datasets of public health importance with the scientific and policy community.
- Support the building of data management capacity for the Network by providing course content and placements for students and trainees in research data management.

The ISHARE2 project is pursuing these objectives in innovative ways:

- The 'Centre-in-a-Box' (CiB) is a research data management appliance that hosts snapshots (analytical databases) of the HDSS's operational database and provides a standard open source data extraction, transformation and loading utility (Pentaho Kettle[4]), a data documentation tool (Nesstar Publisher[5]) to generate DDI (Data Documentation Initiative[6]) metadata. The different software components are hosted in a free for use virtual environment (VMWare ESXi[7]) for ease of maintenance and support. Even though INDEPTH member centres use a variety of different database and data processing environments, the CiB facilitates standardised training, data transformation, and data quality measures to

[2] Walport M, Brest P. Sharing research data to improve public health. Lancet. 2011;377(9765):537-9
[3] Sankoh O, Ijsselmuiden C. Sharing research data to improve public health: a perspective from the global south. Lancet. 2011;378(9789):401-2.
[4] http://community.pentaho.com/
[5] http://www.nesstar.com/software/publisher.html
[6] http://en.wikipedia.org/wiki/Data_Documentation_Initiative
[7] http://en.wikipedia.org/wiki/VMware_ESXi#VMware_ESXi

provide harmonised analytical datasets suitable for multi-site analyses. The CiB reduces the learning curve for data managers to master the often involved process of extracting analytical datasets from complex relational databases to a standard environment with a few well-chosen tools that eases the training and support required to produce standardised and documented data sets across multiple sites.

- The INDEPTH Data Repository is a web-based repository (The World Bank's NADA[8]) to host research datasets and to serve data documentation to data users using the internationally accepted DDI Codebook metadata standard[9].
- INDEPTH (prefix 10.7796) is registered with DataCite[10] through GESIS - Leibniz Institute for Social Science to allocate digital object identifiers for all the data sets on the INDEPTH Data Repository to facilitate the citation of datasets served on the repository.

Available Data on the INDEPTH Data Repository

The INDEPTH Data Repository is a long term project of the Network and the datasets available in the repository will continue to expand in concert with the Network's effort to build its research data management capacity. At the launch of the Repository on 1st July 2013, the repository will contain the detail datasets underlying the indicators on INDEPTHStats for eight Network Centres and the data from an INDEPTH collaboration on the study of the epidemiology of epilepsy in demographic sites (SEEDS).

The detail datasets will contain data in event-history format for approximately 800,000 individuals representing more than 3.7 million person years of observation. The dataset format corresponds to the standard micro-dataset format recently published by INDEPTH[11] and contains a data record for each observed individual demographic event. The following events are recorded:

Event	Description
Birth	The birth of an individual to a resident female
Enumeration	Starting event for all individuals present at the baseline census of the surveillance area. It is the date on which the individual was first observed to be present in the surveillance area during the baseline census.
In-migration	The event of migrating into the surveillance area.
Out-migration	The event of migrating out of the surveillance area.
Location exit	The event of leaving a residential location within the surveillance area to take up residence in another residential location within the surveillance area.
Location entry	The event of taking up residence in a residential location within the surveillance area following a location exit event. Note that location exit and entry are actually two parts of the same action of changing residential location and as such happen on the same event date.
Death	The death of the individual under surveillance. The date of death is the

[8] http://www.ihsn.org/home/software/nada
[9] http://www.ddialliance.org/Specification/
[10] http://datacite.org/
[11] Sankoh O, Byass P. The INDEPTH Network: filling vital gaps in global epidemiology. Int J Epidemiol. 2012;41(3):579-88.

	event date.
Delivery	The event of a pregnancy end after 28 weeks of gestation, which may or may not result in the birth of one or more individuals (represented in this dataset by a BTH event linked to this delivery event)
Observation end	An event inserted when a data set is right censored at an arbitrary date and this individual remained under surveillance beyond this date. The right censor date is the date of this event.
Last observation	An event indicating the last point in time on which this individual was observed to be present and under surveillance. Event date equals observation date in this instance. Normally there should be no individuals with this event as their last event if the right censoring date is prior to the start of the last complete census round.

Table 1: Events

Each record in the dataset contains the following attributes:

Attribute	Description
Record number	A sequential number uniquely identifying each record in the data file
Centre identifier	An identifier issued by INDEPTH to each member centre
Individual identifier	A number uniquely identifying all the records belonging to a specific individual in the data file. For data anonymization purposes, this number is not the same as the identifier used by a contributing centre to identify the individual, but the contributing centre should retain a mapping from this identifier to their identifier
Country identifier	ISO 3166-1 numeric code of the country in which the surveillance site is Situated
Location identifier	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. For data anonymization purposes, this identifier is not the same as the identifier used internally by the contributing centre, but the contributing centre should retain a mapping of this identifier to their internal location identifier
Date of birth	The date of birth of the individual
Event	A code identifying the type of event that has occurred (Table 1)
Event date	The date on which the event occurred
Observation date	Date on which the event was observed (recorded), also known as surveillance visit date
Event count	The total number of events associated with this individual in this data set
Event number	A number increasing from 1 to event count for each event record in order of event occurrence

Table 2: Record layout

Far more detailed documentation of the available datasets can be found on the INDEPTH Data Repository itself.

B. INDEPTH Population Statistics (INDEPTHStats)

Why INDEPTHStats?

Throughout most of low- and middle-income countries, billions of people are born and die without any registration; their lives go unrecorded and are unable to influence health and social policies and programming. But this is not true of over 3 million people who live in the 48 populations covered by the health and demographic surveillance systems (HDSSs) that are members of the INDEPTH Network, run by 40 research centres in 20 countries in Africa, Asia and Oceania. Making their lives count for health and social policy has been the aim of the INDEPTH Network since its start in 1998. Now, summary statistics from these HDSSs will be much more widely available through INDEPTHStats, a website developed by the INDEPTH Network for visualising key demographic indicators.

INDEPTHStats displays the longitudinal health and demographic results generated from the INDEPTH member centres in Africa, Asia and Oceania.

INDEPTHStats is freely available to everyone, and will provide researchers, government officials and policymakers, among others, health and demographic information that can guide their decision-making, including crude birth and death rates, age-specific fertility and death rates, infant, child, and under five mortality rates, as well as numerous other health and demographic indicators. Additional indicators, such as death rates by cause of death, will be added in the near future. The indices can be displayed either by single centre over time, or across multiple centres.

New data will be added annually on 1st July each year.

Making these data available to everyone has been made possible by funding from the Hewlett Foundation, Sida/Research Cooperation Unit, Wellcome Trust, Rockefeller Foundation and the Gates Foundation combined with the generous, free provision of the data by the member HDSSs. All the data have been subjected to rigorous technical checks, first at the individual HDSSs and then within INDEPTH.

INDEPTH looks forward to receiving comments from data users, and suggestions for how to further enhance the value provided by this new source of critical health and demographic data for research, policy making and health and social programming.

Available indicators on INDEPTHStats

INDEPTHStats will display the following five groups of indicators based on HDSS data: population, fertility, migration, mortality and cause of death. We intend to add more groups of indicators in the near future.

1. HDSS Population Data

The available indicators in this group are sex ratio, proportion <15, proportion 15-49, and proportion 60+. We provide brief definitions below.

Population

Population is the person-years, i.e., the sum of years or fraction of years lived in the HDSS by the residents of the HDSS during a given period or calendar year. All rates are computed using person-years as denominator.
Minimum time to be considered a resident varies from one HDSS to the other. On average this is six months.
Population is NOT the number of residents in the HDSS at a given date (beginning, mid, or end of the year).

Sex ratio

Sex ratio is the ratio of males to females, i.e. the number of HDSS male residents divided by the number of HDSS female residents, expressed in number of males for 100 females.

Proportion <15

It is the number person-years lived before the 15th birthday divided by the total number of person-years.

Proportion 15-49

It is the number person-years lived between the 15th and 60th birthdays divided by the total number of person-years.

Proportion 60+

It is the number person-years lived after the 60th birthday divided by the total number of person-years.

2. HDSS Fertility Data

The indicators available are crude birth rate (CBR), Age-specific fertility rates (ASFR), Total fertility rate (TFR), Mean age at childbearing (MAC) and Sex ratio at birth (SRB)

Fertility

Fertility is measured using the number of births to HDSS female residents aged 15 to 49 .

Crude birth rate (CBR)

CBR is the number of births divided by the person-years of all ages, expressed in births per 1000 person-years.
NB: CBR depends on age-structure, which evolves from one calendar year to the next and from one HDSS to the other. Therefore CBRs are not strictly comparable over the period and across HDSS.

Age-specific fertility rates (ASFR)

ASFR is the number of births in a specific age-group divided by the person-years lived in that age-group, expressed in births per 1000 people. The seven age-groups are 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49.

Total fertility rate (TFR)

TFR is the sum of ASFR multiply by 5. It is interpreted as the mean number of births that a hypothetical cohort of women would have had by experiencing the fertility conditions of a particular calendar year over its lifetime.

Mean age at childbearing (MAC)

MAC is computed as the sum of age-specific fertility rates (ASFR) weighted by the distribution of females at the mid-point of each age group as per the life-table estimate, divided by the sum of the age-specific rates. It is interpreted as the mean age of mothers at the birth of their children if women were subject throughout their lives to the age-specific fertility rates observed in a given year.

Please note that the MAC computed for INDEPTHStats is **not** the standard MAC as found in textbooks, since it is weighted using the distribution of women by age as per the life-table estimate.

Sex ratio at birth (SRB)

SRB is the number of male births divided by the number of female births, expressed in male births per 100 female births.

3. HDSS Migration Data

The indicators available are Crude in-migration rate (CIMR), Crude out-migration rate (COMR), Age-specific in-migration rates (ASIMR), Age-specific out-migration rates (ASOMR), Crude net-migration rate (CNMR) and Crude gross-migration rate (CGMR).

Migration

Migration is defined as a change of residence. Minimum time to be considered a resident varies from one HDSS to the other (see Population). In-migrants are residents after they have spent more than that minimum time in the demographic surveillance area. Out-migrants are no longer residents after they have spent more than that minimum time outside the demographic surveillance area.

Crude in-migration rate (CIMR)

CIMR is the number of in-migrations divided by the person-years of all ages, expressed in in-migrations per 1000 person-years.

NB: CIMR depends on age-structure, which evolves from one calendar year to the next and from one HDSS to the other. Therefore CIMRs are not strictly comparable over the period and across HDSS.

Crude out-migration rate (COMR)

COMR is the number of out-migrations divided by the person-years of all ages, expressed in out-migrations per 1000 person-years.

NB: COMR depends on age-structure, which evolves from one calendar year to the next and from one HDSS to the other. Therefore COMRs are not strictly comparable over the period and across HDSS.

Age-specific in-migration rates (ASIMR)

ASIMR is the number of in-migrations in a specific age-group divided by the person-years lived in that age-group, expressed in in-migrations per 1000 person-years.

Age-specific out-migration rates (ASOMR)

ASOMR is the number of out-migrations in a specific age-group divided by the person-years lived in that age-group, expressed in out-migrations per 1000 person-years.

Crude net-migration rate (CNMR)

CNMR is the number of in-migrations minus out-migrations divided by the person-years of all ages, expressed in net-migrations per 1000 person-years. CNMR may be positive or negative.
NB: CNMR depends on age-structure, which evolves from one calendar year to the next and from one HDSS to the other. Therefore CNMRs are not strictly comparable over the period and across HDSS.

Crude gross-migration rate (CGMR)

CGMR is the number of in-migrations plus out-migrations divided by the person-years of all ages, expressed in migrations per 1000 person-years.
NB: CGMR depends on age-structure, which evolves from one calendar year to the next and from one HDSS to the other. Therefore CGMRs are not strictly comparable over the period and across HDSS.

4. HDSS Mortality Data

Mortality indicators available are Crude death rate (CDR), Age-standardised crude death rate (ASCDR), Age-specific mortality rates (ASMR), Neonatal mortality rate per 1000 person-years, Neonatal mortality rate per 1000 live births, Infant mortality rate 1q0, Child mortality rate 4q1, Under-five mortality rate 5q0, Adult mortality rate 45q15 and Life expectancy at birth e0.

Mortality

Mortality is measured using the number of deaths of HDSS residents.

Crude death rate (CDR)

CDR is the number of deaths divided by the person-years of all ages, expressed in deaths per 1000 person-years.
NB: CDR depends on age-structure, which evolves from one calendar year to the next and from one HDSS to the other. Therefore CDRs are not strictly comparable over the period and across HDSS.

Age-standardised crude death rate (ASCDR)

ASCDR is a weighted average of the age-specific mortality rates, expressed in deaths per 1000 person-years. The weights are the proportions of person-years in the corresponding age over the entire period of HDSS observation.
NB: ASCDR depends on the chosen age-structure, which is different from one HDSS to the other. However ASCDRs are comparable over the calendar years of observation for the same HDSS.

Age-specific mortality rates (ASMR)

ASMR is the number of deaths in a specific age-group divided by the person-years lived in that age-group, expressed in deaths per 1000 person years.

Neonatal mortality rate per 1000 person-years

Neonatal mortality rate is the number of deaths in the first 28 completed days of life divided by the number of person-years lived in the HDSS within the first 28 days, expressed in deaths per 1000 person-years.

Neonatal mortality rate per 1000 live births

Neonatal mortality rate is the standard rate obtained by dividing the number of deaths in the first 28 completed days of life by the number of live births in the HDSS, expressed in deaths per 1000 live births.
NB: This indicator is slightly biased depending on migrations in and out of the HDSS between birth and 29th day. This is because infants migrating in the HDSS (becoming new residents) after their birth and before their 29th day are counted in the number of deaths if they die, but are not counted as births. Also, infant migrating out of the HDSS before their 29th day are counted among living births, but are not counted as deaths if they die after out-migrating and before their 29th day.

Infant mortality rate 1q0

Infant mortality rate is the life-table (Kaplan-Meier) estimate using exact dates of death and of censoring. It is expressed in deaths per 1000 live births and interpreted as the probability to die before 1st birthday.
NB: This indicator is slightly biased depending on migrations in and out of the HDSS between birth and 1st birthday. This is because infants who in-migrated are accounted for if they spent at least the minimum time required to be a resident in the HDSS (if they die during this time span, they are not considered residents), whereas infants who out-migrated are accounted for the total time spent in the HDSS from birth (or age at in-migration) to age at out-migration.

Child mortality rate 4q1

Child mortality rate is the life-table (Kaplan-Meier) estimate using exact dates of death and of censoring. It is expressed in deaths per 1000 person-years surviving their 1st birthday and interpreted as the probability to die between 1st birthday and 5th birthday.
NB: This indicator is slightly biased depending on migrations in and out of the HDSS between 1st birthday and 5th birthday. This is because children who in-migrated are accounted for if they spent at least the minimum time required to be a resident in the HDSS (if they die during this time span, they are not considered residents), whereas children who out-migrated are accounted for the total time spent in the HDSS from 1st birthday (or age at in-migration) to age at out-migration.

Under-five mortality rate 5q0

Under-five mortality rate is the life-table (Kaplan-Meier) estimate using exact dates of death and of censoring. It is expressed in deaths per 1000 live births and interpreted as the probability to die before 5th birthday.

NB: This indicator is slightly biased depending on migrations in and out of the HDSS between birth and 5th birthday. This is because children who in-migrated are accounted for if they spent at least the minimum time required to be a resident in the HDSS (if they die during this time span, they are not considered residents), whereas children who out-migrated are accounted for the total time spent in the HDSS from birth (or age at in-migration) to age at out-migration.

Adult mortality rate 45q15

Adult mortality rate is the life-table (Kaplan-Meier) estimate using exact dates of death and of censoring. It is expressed in deaths per 1000 people surviving their 15th birthday and interpreted as the probability to die between 15th birthday and 60th birthday.

NB: This indicator is slightly biased depending on migrations in and out of the HDSS between 15th birthday and 60th birthday. This is because adults who in-migrated are accounted for if they spent at least the minimum time required to be a resident in the HDSS (if they die during this time span, they are not considered residents), whereas adults who out-migrated are accounted for the total time spent in the HDSS from 15th birthday (or age at in-migration) to age at out-migration.

Life expectancy at birth e0

Life expectancy is the life-table (Kaplan-Meier) estimate using exact dates of death and of censoring. It is expressed as an average number of years that a hypothetical cohort would have lived by experiencing the health conditions of a particular calendar year over its lifetime.

NB: This indicator is slightly biased depending on migrations in and out of the HDSS. This is because people who in-migrated are accounted for if they spent at least the minimum time required to be a resident in the HDSS (if they die during this time span, they are not considered residents), whereas people who out-migrated are accounted for the total time spent in the HDSS from 1st birthday (or age at in-migration) to age at out-migration.

5. HDSS Cause of Death Categories

Cause of death is determined by administering verbal autopsies (VAs) and analysing them using InterVA-4. The data on cause of death from HDSSs are available in the following broad categories:

- VAs-01 – Infectious and parasitic diseases
- VAs-02 – Neoplasms
- VAs-03 – Nutritional and endocrine disorders
- VAs-04 – Diseases of the circulatory system
- VAs-05 – Respiratory diseases
- VAs-06 – Gastrointestinal disorders

- VAs-07 – Renal disorders
- VAs-08 – Mental and nervous system disorders
- VAs-09 – Pregnancy, childbirth and puerperium-related disorders
- VAs-10 – Neonatal causes of death
- VAs-11 – Stillbirths
- VAs-12 – External causes of death
- VAs-98 – Other and unspecified non-communicable disease
- VAs-99 – Cause of death unspecified or unknown