BMZ Federal Ministry
for Economic Cooperation
and Development

Future-Makers.
Building the future.
Let's join forces.

Collection, use and assessment of secondary and primary data for results measurement in health

A Guide for Practitioners

Published by giz In cooperation with KFW

Content

List of abbreviations

AIS	AIDS indicator survey
AIDS	Acquired immunodeficiency syndrome
DHS	Demographic and health survey
DFID	Department for International Development
EPI	Expanded Programme of Immunization
GDC	German development cooperation
GIZ	Deutsche Gesellschaft für Internationale Zusammenarbeit
GPS	Geographical Positioning System
IHP+	International Health Partnership and related initiatives
IHSN	International Household Survey Network
HH	Household
HMIS	Health management information system
HIV	Human Immunodeficiency Virus
LSMS	Living Standard Measurement Surveys
M&E	Monitoring & Evaluation
MIS	Malaria Indicator Survey
MICS	Multi Indicator Cluster Survey
NGO	Non-governmental Organisation
PMTCT	Prevention of Mother-to-Child Transmission
PDA	Personal Digital Assistant
SAM	Service availability mapping
SPA	Service Provision Assessment
TB	Tuberculosis
UNFPA	United Nations Population Fund
UNGASS	United Nation General Assembly Special Session on HIV/AIDS
UNICEF	United Nations Children's Fund
WHO	World Health Organisation

List of boxes, figures, tables

1. Introduction

The main aim of this manual is to give an overview of available data sources which can be used for results measurement of health programmes and projects as well as guidance on how to assess the quality of indicators retrieved from these different data sources.

The next two chapters introduce the concept of results measurement, an internationally recognized monitoring and evaluation framework, core indicator domains to monitor progress in health (chapter 2), different types of indicators and key quality criteria of data (chapter 3). Readers familiar with results measurement and the IHP+ framework for monitoring and evaluation of health system strengthening can skip both chapters and directly refer to chapter 4 which provides an overview of sources for data collection with corresponding hyperlinks. Data sources included in this manual are censuses, a variety of population-based and health facility surveys, routine health service data and vital registration systems (see chapter 4.1). Quality issues and key epidemiological concepts important to assess the quality of data are described in chapter 4.2-4.4. Chapter 5 guides through some epidemiological aspects to consider when planning for complementary generation of primary data where the secondary data sources give too little and to unspecific information. Some notes on data in emergency settings are given in chapter 6.

The manual is based on available literature, practical experience and expertise and previous work into results measurement and impact evaluation in health programmes supported by German development cooperation (GDC)[1].

The manual aims to help programmes and projects particularly in the planning phase but also during project implementation as well as to revise the conceptual framework and indicator sets for monitoring projects.

[1] See "Wirkungsanalyse und Wirkungsmessung in Gesundheitsvorhaben der deutschen Entwicklungszusammenarbeit, Meth oden, Situationsanalyse und Empfehlungen" June 2011 and the GIZ publication "Baselineerhebung. Ein Leitfaden zur Planung, Durchführung, Auswertung und Nutzung der Ergebnisse".

2. Results measurement in health

Results measurement in programmes and projects of the GDC aims at providing evidence how well a project/programme performs and whether it reaches its targets in an overall health context. The main objective of measuring results is to provide evidence on the extent to which a project/programme achieves its targets, goals and objectives. Results measurement should help to revise or refine strategies. Moreover, monitoring and evaluation of projects or programmes has become increasingly important to justify efforts and to ensure that resources are spent efficiently.

There continues to be a need to harmonize and align results measurement with international result measurement standards and country monitoring frameworks. Country monitoring frameworks are increasingly based on frameworks formed by MDG monitoring or developed as part of the work of the International Health Partnership (IHP+). The improved availability of data and indicators even for sub-national regions in sources openly available in the worldwide net provides great opportunities for the usage of comparable indicators and the establishment of joint evaluation processes.

Results measurements should include the micro, meso and macro levels in health systems, thus the local or implementation level, organizations and structures as well as the national level including policy and policy dialogue. Indicators best include immediate, intermediate and, if possible, long-term changes. Long-term changes are measured using impact indicators such as maternal and infant mortality. Intermediate indicators reflect changes within a shorter period such as readiness and coverage or output and outcome indicators. Immediate changes are measured by input indicators. Results measurement uses frameworks through which inputs and processes lead to outputs, outcomes and final impact improvements (see figure 1).

Results measurement can be understood rather as a practical monitoring exercise than a methodological rigorous evaluation and is not to be confounded with rigorous impact evaluation. Impact evaluations require a counterfactual analysis and aim at quantifying the exact impact an intervention has on health outcomes such as child mortality, maternal mortality or HIV-prevalence and respective proxy indicators[2].

Result measurement thus accepts that the exact attribution of efforts of health projects and programmes with regard to health impacts is not possible in most situations. Still, results measurements can inform about important advances in access to services, population coverage of interventions as well as changes in knowledge, attitude and practise. If several indicators on the output and outcome level are improving in response to measurable and sustained efforts at the input and process level of a health programme one may carefully suggest that improvements in outcome or impact indicators might be due to the implemented intervention. Stronger evidence can be derived when progress is also measured in a comparison area to exclude that improvements are not due to temporal changes (such as general improvement in health). However, these so called ‘plausibility' approaches demand careful assessment whether the intervention / the project has been mainly responsible for improvements, or possibly other interventions implemented at the same time (for a more detailed definition refer to the box 1 in the annex) [2].

WHO has suggested a common monitoring and evaluation framework to serve for all monitoring activities in a country. This framework envisages the alignment to country systems as part of the Paris declaration. It encourages donors and global health initiatives to join in with the aim of a single integrated monitoring and evaluation system. Such a common system based on various data sources and including indicators along the six WHO building blocks would make it possible to identify gaps but also to provide a more complete picture of overall progress [5].

[2] More also on the debate on impact evaluation in The homepage of the **International Initiative for Impact evaluation** (3ie) offers good working papers on opportunities for rigorous but feasible impact evaluation. *http://www.3ieimpact.org/admin/pdfs_papers/11.pdf*

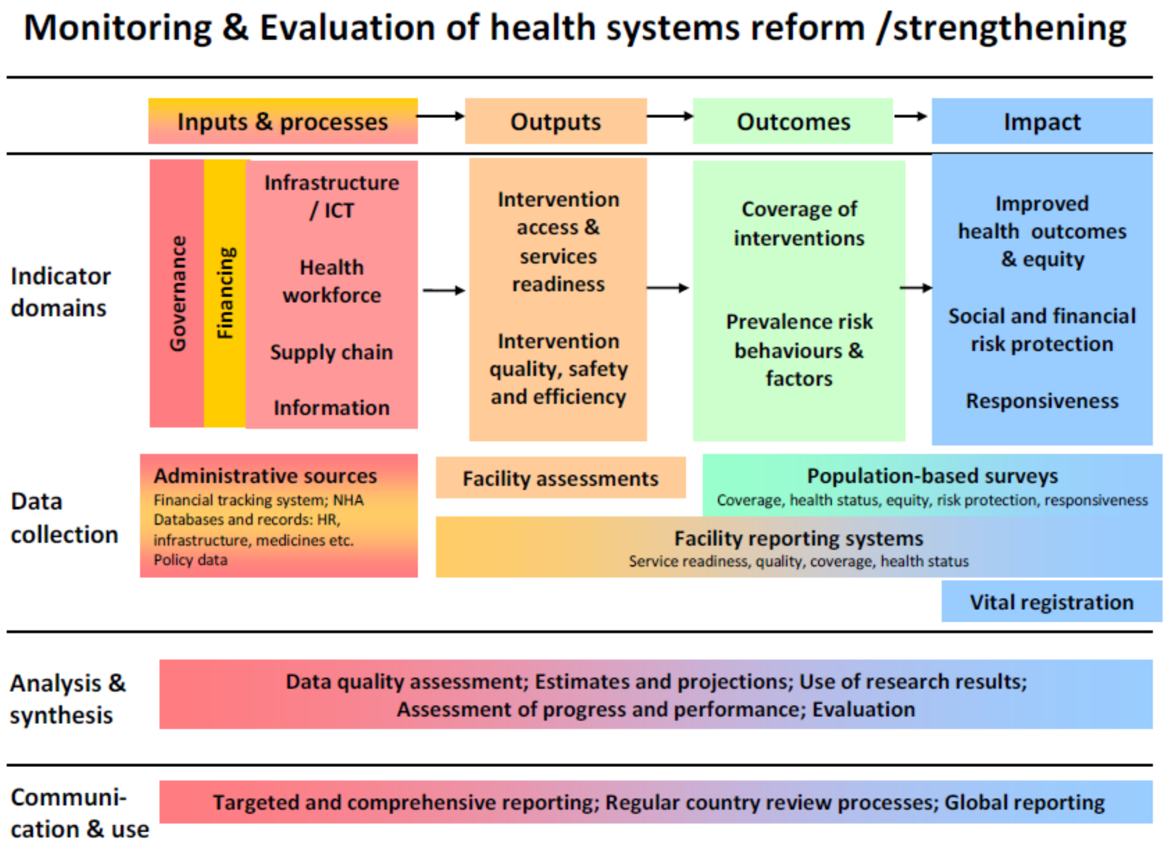

Monitoring & Evaluation of health systems reform /strengthening
Inputs & processes
Outputs
Outcomes
Impact
Indicator domains
Governance
Financing
Infrastructure / ICT
Health workforce
Supply chain
Information
Intervention access & services readiness
Intervention quality, safety and efficiency
Coverage of interventions
Prevalence risk behaviours & factors
Improved health outcomes & equity
Social and financial risk protection
Responsiveness
Data collection
Administrative sources
Financial tracking system; NHA
Databases and records: HR, infrastructure, medicines etc.
Policy data
Facility assessments
Population-based surveys
Coverage, health status, equity, risk protection, responsiveness
Facility reporting systems
Service readiness, quality, coverage, health status
Vital registration
Analysis & synthesis
Data quality assessment; Estimates and projections; Use of research results; Assessment of progress and performance; Evaluation
Communi-cation & use
Targeted and comprehensive reporting; Regular country review processes; Global reporting

Fig 1: Framework for monitoring and evaluation of health system strengthening [5]

In addition to the overall framework, for many of the indicator domains separate 'sub-frameworks' have been suggested such as for example, for the evaluation of governance [6], or the evaluation of country health information systems [7, 8]. These frameworks greatly assist in planning and structuring monitoring and evaluation. They also assist in examining the flow of effects.

The framework published by WHO shown above also provides an overview of which data and indicators can be obtained from which sources. **Outcome indicators** are part of the population-based surveys (DHS, AIS, MICS, behavioural surveys) which also partly cover **impact indicators** such as improved health outcomes (infant, child or maternal mortality and equity), but also information on social protection. In high income countries as well as some middle income countries data on mortality are derived from vital registration systems. **Output indicators** include data on access, quality, efficiency and utilization and are mainly available from health facility surveys or routine facility reporting systems (mostly referred to as health management information systems HMIS). **Input indicators** include elements of what is also described by WHO as the 'health system building blocks' [9]. Input and process indicators are available from administrative records (health provider records and registers, national health accounts) and project and programme documentation.

Thus most importantly, the proposed framework encourages to measure progress along the line from input to outcome to impact indicators. It encourages to use different sources and to use standard definitions as much as possible, so that available data from country sources can be used.

3. Types of data and key concepts in data quality

Result measurement can be based on primary and secondary data. **Primary data** are data collected by the researcher or user (such as the project, implementer or project evaluator). **Secondary data** is data collected by someone other than the user. Here, the sources presented under 3 a. are the most relevant. But also qualitative information collected by others or found in published articles may be used.

Secondary data analysis can save time and costs and free resources for additional complementary data collection, for example to explain findings or detailed analyses of certain key research questions.

Methods to collect information are divided in two main groups: quantitative and qualitative methods. Quantitative data are primarily used to get information that relates to quantity and provides answers to questions such as: "How much?", "How many?", "How frequently?"

Quantitative data are obtained from

- Representative surveys such as sample household surveys, census, or surveys in special target groups such as the DHS, MICS surveys as discussed in 4.1. & 4.2.

- Facility surveys, here health facility surveys or census as discussed in 4.1.

- Administrative data/routine data collection such as data collected through the national health information system (HMIS) as discussed in 4.3.

- Geographical positioning systems (GPS) data which help to locate people or institutions and thereby answer the question: "Where?"

  Qualitative data try to answer questions like "Why?" and "How?" These data provide information on the population's (or sub-groups) perceptions, values, norms, concepts and experiences in their socio-cultural context. Results from qualitative research will not present a 'count' but explanations and facilitate deeper understanding of behaviors. Qualitative studies are particularly useful to prepare for quantitative data collection in order to formulate relevant questions and to include all possible answer options. Qualitative data are also important to explain findings from quantitative research in more depth. Thus qualitative and qualitative data should be seen as complementary and their combination is nowadays considered standard.

Qualitative data are obtained through the following instruments

- Questionnaires interviews (largely using unstructured, or semi-structured interview guidelines with open-ended questions

- Key informant interviews (mostly using semi-unstructured interview guidelines)

- Group interviews (e.g. focus group discussions)

- Projective techniques (e.g. open ended stories)

- Critical document(ary) analysis

Analysis of quantitative data

Quantitative data are in general analyzed presenting counts and frequencies. Information might be disaggregated in relation to 'background characteristics' being age, geographical region, education or wealth quintile (stratified analysis). The latter analysis gives indications on the level of equity. Often presentation is done using frequency tables (see DHS).

Further, quantitative data can be used to derive measures of impact. Effect measures include odds ratios, risk ratios or rate ratios.

Odds, incidence risk, incidence rates, odd ratio, risk ratio and rate ratio

Odds are calculated by dividing the number of people with the outcome by the number of people without the outcome. In contrast incidence risks are derived by calculating the number of people with the outcome by all people included in the study. Incidence risks can be calculated from cohort studies were a group of people are followed for a specified time. In contrast odds are derived from case-control studies.

The odds ratio of an outcome compares the odds of having a particular outcome in a population exposed to a risk factor with the odds in a population not exposed. Risk ratios compare the incidence risk of an outcome between two populations. An example is that the incidence of traffic acci

dents is two times higher in population A than population B. It measures the magnitude of an exposure effect on the incidence. Rate ratios are based on incidence rates. Incidence rates measure the incidence based on person-time at risk. Incidence rates are used for common outcomes. And the measurement takes into consideration that people experience different time lengths of exposure. An example can be the incidence of tuberculosis in miners where one person might work as a miner for 4 months, another 4 years. The rate ratio of the incidence of tuberculosis, comparing miners with people who are not miners might be for example five. This means that the chance getting tuberculosis is 5 times higher in miners.

These effect measurements quantify the effect, the variable of interest (i.e. independent or explaining variable) or the intervention has on the final outcome of interest (which can be output, outcome or impact indicators). Whether the intervention is significantly more effective to improve health outcomes (or increase coverage or knowledge) than no or an alternative intervention can be assessed using chi-squared test, t-tests, regression methods or other statistical tests employing significance testing.

Different study types are used to obtain effect measurements, such as case-control studies, cohort studies or experimental studies using a randomized or non-randomized controlled study design[3].

Most relevant for impact evaluations of health programmes are cluster-randomized trials or pre-and post-survey design with concurrent comparison group/area. Both these study designs use not individuals but communities, or other geographical areas (e.g. districts) as study unit. This should be distinguished from conventional medical studies looking for example at the effect of smoking on health where the unit of analysis is the individual.

[3] A good introduction can be found in: Health Care Evaluation (Understanding Public Health) [Paperback]. Sarah Smith, Don Sinclair, Rosalind Raine, Barnaby Reeves. 2005. Open University Press. *www.openup.co.uk*. ISBN -10: 0335218490

Both, cluster-randomized trials and the 'plausibility approach' using pre-and post-survey design with concurrent comparison group/area are being used increasingly to evaluate the effect of health strategies to improve health (see more examples in chapter 7).

Information and summary estimates from trials are regularly compiled in Cochrane reviews [4]. However, as outlined before, whereas impact evaluation is based on rigorous study designs, results measurements is rather a practical concept which tries to assess progress without measuring the exact effect. Thus results measurement does not include significance testing but only stratification of results (frequency tables in relation to background characteristics)

Analysis of qualitative data

Qualitative data are analysed using an inductive or deductive methods, and most often a combined approach. The initial grouping of data is commonly based on the key themes reflected (for example) in the semi-structured interview guideline (deductive). In a second round – or parallel – data are explored with a view on additional themes and issues evolving from the data (inductive). By this, content analysis is applied and key analytic categories and themes are identified using coding schemes which are subsequently developed. The interpretation includes looking for patterns of how people understand and explain perceptions, experiences and behaviour. Preparation for data analysis may include the full transcription of the interviews if they were tape-recorded or comprehensive summaries of interview notes. Analysis is often done manually; there are also software programmes for coding such as NVIVO®.

Quality of data

Validity (accuracy), reliability, precision, and completeness are defined as key quality criteria of data. Other quality criteria include the timely availability of information. In the recent years also 'integrity' of data has been added as a quality criterion. Integrity includes that data manipulation for scientific or political reason is excluded. Also confidentiality of the respondents needs to be assured. Box 2 in the annex lists various dimensions of key quality concepts with operational definitions and examples.

[4] *http://www.thecochranelibrary.com/view/0/index.html*

4. Secondary data: Retrieval and use

4.1. Available data sources

Population-based surveys

Demographic and Health Surveys (DHS), AIDS or Malaria Indicator surveys (AID & MIS) by Macro International

DHS surveys are being done since 1984 and built on previous experience from World Fertility Surveys. Standard DHS Surveys have large sample sizes (usually between 5,000 and 30,000 households) whereas Interim DHS Surveys focus on coverage and does not include mortality data. They have smaller sample sizes (2,000–3,000 households). A modular questionnaire is used. The main modules include questions on fertility, fertility preferences and family planning, child and maternal health, wealth, education and other variables. Since the mid 90ths modules to assess adult and maternal mortality have been added to many DHS. Today different modules are available, and both women and men are commonly interviewed. Many key indicators used for the assessment of progress in health are available from DHS data sets, such as child and infant mortality, total fertility and many others. DHS data are commonly rated as of good quality. National estimates have a high precision but sub-regional or sub-group estimates are less precise (see below more on sample size) and point estimates should be used with certain caution.

Full country reports are released and made publically available approximately one year after the survey has been conducted on the homepage[5] . The full data sets are also made available for research purposes, but registration is required with mention of what the research project is intended for.

In response to the need to monitor national HIV/AIDS or malaria programmes effectively, AIDS and Malaria Indicator surveys (AIS, MIS) were introduced by Measure using a standardised tool.

Periodicity: Every 3 to 5 years; Sample size: 5000 to 30000 households (hh); Main indicator fields: Fertility (e.g. Total fertility rate, adolescent birth rate), family planning (e.g. contraceptive prevalence, unmet need), maternal and child mortality (e.g. newborn, infant child mortality rate and maternal mortality ratio), maternal and child health (e.g. antenatal care attendance, skilled attendance at birth) The **HIV/AIDS Indicator Survey (AIS)** monitoring tool provides information which is required for President's Emergency Plan, UNGASS, but also other indicators while ensuring comparability of findings across countries and over time. The AIS commonly includes modules to assess HIV/AIDS knowledge, attitudes, and behaviour as well as HIV Prevalence. Periodicity: Every 3 to 5 years; Sample Size: 5000-10000 hh; Main indicator fields: HIV/AIDS, knowledge (e.g. correct knowledge among youth), attitude, behaviour, HIV prevalence in relation to education and wealth index, collection of UNGASS indicators

The **Malaria Indicator Survey (MIS)** provides much needed information on the use of mosquito nets, prevention of malaria during pregnancy, prompt and effective treatment of fever in young children, and indoor residual spraying of insecticide to kill mosquitoes. The MIS was developed by the Monitoring and Evaluation Working Group (MERG) of the Roll Back Malaria initiative. Periodicity: Every 3 to 5 years or combined; Sample size: 4000 to 10000 hh; Main indicator fields: Prevention (bednets, intermittent preventive treatment), use of health care, anemia and malaria prevalence

[5] see *www.measuredhs.com/* and click on 'Where we work' and then a list of countries is available. Subsequently all surveys ever done in the respective country are listed and it is possible to immediately download all reports.

Multiple Indicator Cluster Surveys (MICS) by UNICEF

Multiple Indicator Cluster Surveys have been conducted since the mid-1990s by UNICEF. The first survey round was developed in response to the World Summit for Children and included 60 countries with the aim to produce statistically sound and internationally comparable estimates of a range of indicators in the areas of health, education, child protection and HIV/AIDS. Today the 4th round of MICS surveys (2009-2011) are ongoing and a few reports have already been published. Harmonization of survey questions and modules with DHS was done in the recent years and the new rounds also include a questionnaire directed to men.

MICS usually include around 10,000 to 20,000 households. Sampling methods differ little from DHS and estimates are broadly comparable with DHS data.

The MICS data are also publically available at the Unicef.org statistics website or at http://www.childinfo.org/ but the homepage is a bit less user friendly than the DHS homepage. Full reports describing methodology and main results are available but also tables for each country. Sub-national data or estimations disaggregated for background characteristics (education, wealth and others) are not found in these tables. The quality of the MICS data is good, and estimates have a good precision for the national level estimates.

Periodicity: Every 5 years; Sample size: 5000 to 30000 hh; Main indicator fields: Child mortality (infant and child mortality), nutrition (e.g. underweight, exclusive breastfeeding), vaccination (e.g. coverage with Tb, measles, DPT3 and others), Malaria prevention (e.g. bednet use, intermittent preventive treatment) water and sanitation and others on fertility and maternal and child health as above

Living standard measurement surveys (LSMS) by World Bank

Living standard measurement surveys are supported by the World Bank since the 1980s. They aim to provide information on of unemployment, poverty and health care use and have been carried out in 35 countries. Information on health spending (out-of-pocket expenses is typically derived from these surveys.

Periodicity: very, repeated since 1980; Sample size: 1000 to 10000 hh; Main indicator fields: Assets, expenditure on various issues (including out-of-pocket expenditure), mortality, utilization of health services, education, agriculture

Household Budget Surveys

Many countries, foremost high income, but also several low and middle income countries have conducted household budget surveys which include economic information such as health spending but also distance to health facility.

Periodicity: About every 5 years; Sample size: 10000-20000 hh; Main indicator fields: Poverty and living standards (assets, expenditure on education, health, food and others), access to education and health

World Health Surveys by WHO

The survey includes data information on health, access to health and health spending and has been carried out in over 72 high, middle and low income countries.

Periodicity: Once conducted in 2002-2003; Sample size: 1000 to 10000 hh, Main indicator fields: Adults and child mortality, morbitiy, use of health services, health care expenditure (e.g. out-of-pocket measurements, insurance coverage and costs), preventive measures (malaria)

Vital registration / Census data

Census data primarily give information on population size, population growth, a few economic indicators as well as total counts of birth and deaths. Some censuses have added an additional module to estimate maternal mortality[6]. This has been much promoted as DHS estimates on MMR suffer from wide confidence intervals and sub-national estimates cannot be calculated as sample sizes are too small.

Census data are sometimes available on national websites, but can also be accessed on http://www.internationalsurveynetwork.org/home/

Child Mortality Website

The web-site is owned by the Inter-Agency group of Child Mortality estimation. Updates with new estimations are available once a year, mid September, including a full report on methodological issues. Sources of data are national vital registration systems, data from national population census and/or data collected via household surveys. Methodological papers are also available on this homepage concerning the credibility of the decline of child mortality found in several DHS.

[6] United Nations, Principles and Recommendations for Population and Housing Censuses, Revision 2. 2008, UN, Series M. No 67/Rev2. New York

Birth and death registration – Information from demographic sentinel sites through Indepth Network

Birth and death registration – Information from the Sample Vital Registration with Verbal Autopsy (SAVVY)

Countries without any continuous birth and death registration are increasingly using demographic sentinel sites to obtain in particular data on causes of mortality. These demographic sentinel sites are also used in many countries as a 'laboratory' to test new interventions. One of the most renowned sentinel sites is the Matlab demographic site in Bangladesh, which is operating since the 1970s. However, information from these demographic sentinel sites might suffer from the lack of 'generalisability' as they often include only a small population and results do not necessarily reflect fertility and mortality in the respective country but only the respective area.

To have national representative data from demographic sentinel sites, a surveillance system called 'Sample Vital Registration with Verbal Autopsy (SAVVY)' is now introduced in several low or middle countries without national birth and death registration. Here vital events (birth and deaths) are monitored in a similar way as in demographic sites and verbal autopsy for investigation nto causes of death is done. The system includes several areas in a country so that the information collected provides nationally representative information about levels and causes of mortality as well as other indicators not available from other sources.

Countries with a national birth and death registrations often publish data on causes of deaths annually on the web pages of their respective national statistical institutes. There is no specific web site as yet established to retrieve birth and death data or cause of death from SAVVY systems thus the national institutes supporting the sites need to be contacted.

China and India use demographic sentinel sites since a long time and the Chinese system is rated of high quality and to give representative result.

Routine and administrative data

Some national data come from routine health information systems or separate routine monitoring activities such as those established for vaccination (EPI), tuberculosis, antiretroviral treatment and PMTCT. These separate monitoring systems are to a varying degree integrated into routine health management information system (HMIS), but not everywhere.

UN member states, for example, are obliged to report progress regularly to the UN General Assembly (http://www.unaids.org/en/dataanalysis/monitoringcountryprogress/2010progressreportssubmittedbycountries/).

Indicators included in the UNGASS monitoring are to a large extent based on health services data such as the percentage of adults and children with advanced HIV receiving antiretroviral therapy or percentage of HIV-positive pregnant women receiving antiretroviral treatment to reduce MTCT.

In addition, national health worker registers, drug monitoring registers etc. can be used to get data for certain indicators.

Health facility surveys – Service Provision Assessment
Health facility surveys – Service Availability Mapping
Health facility surveys – Health Facility Census

Three main sources of health facility data are publically available. MEASURE supports the **Service Provision Assessment (SPA)** which is a comprehensive assessment of the capacity of health facilities to offer key services. The assessment concentrates on child health, maternity and newborn care, family planning, sexually transmitted infections (STI/HIV/AIDS) and other infectious diseases and includes infrastructure, equipment and supply, supervision and referral and some information on quality of care assessed as readiness of the health facilities to offer selected services or interventions.

Reports are available together with DHS reports when opening the window available for each country (see above).

Since 2004 WHO has supported health facility assessments called **Service availability mapping (SAM)** in a limited number of countries. These have assessed in particular the human resource situation. At the moment a new questionnaire is being prepared by WHO (SARA).

The third health facility assessment tool is the **Health Facility Census (HFC)** which is supported by the Japanese International Cooperation Agency, which focuses on the physical assets and the status of health facilities. However, as with all censuses, the approach has proved to be very expensive and has only been carried out in a limited number of countries.

Efforts are under way to harmonize assessment tools so that data and indicators can be compared using different surveys [11].

National health accounts

National Health Accounts constitute a systematic, comprehensive and consistent monitoring of resource flows in a country’s health system for a given period and reflect the main functions of health care financing: resource mobilization & allocation, pooling and insurance, purchasing of care and the distribution of benefits. WHO provides guidance for classification to enable cross country comparisons. Data available on the WHO national health account website include much of the indicators used to monitor health financing and social security and include total expenditure on health per capita (exchange rate or PPP), out-of-pockets expenditures and others.

International websites compiling data from above mentioned and other sources

There are several international websites compiling information from many of the above mentioned sources.

WHO Global Health Observatory

Health data from various sources and links to reports (such as MICS) can be found as well as indicator summary tables

International Household Survey Network

Provides an overview of censuses and household surveys ever done in a country.

Gapminder

Compiles data from all countries and different sources. The compilation also includes historical data e.g. from census data 100 years back. Further, the data set include many economic, social and other development indicators. A more elaborated documentation of sources of data is available for selected indicators (go to 'data' and then 'documentation').

e-Atlas of Global Development

A visual guide to the world's greatest challenges and is a practical companion to the World Bank's popular Atlas of Global Development.

Health Systems 20/20 Database

Compiles and analyzes country data and key indicators from multiple sources helping the user to assess the performance of the country´s health system.

A quality criterion of these websites is the availability of so called 'Meta Data'. These should describe the source of the data, whether data were adjusted or any other statistical method has been applied. Also a clear definition of the indicator (nominator/denominator) should be available. Meta data are for example found when consulting the WHO data-base Global Health Observatory Data Repository[7]. Here the indicator definition and where the data come from are indicated if one opens the "i" box. Also other global data bases present meta data. When visiting the indicator list of Gapminder (*www.gapminder.org*) a definition of the indicator and the sources from where the data are obtained is given.

4.2 Quality assessment of census and population-based survey data

The most important quality concepts relevant to population-based surveys are sampling procedures including sample size. A census by definition includes all members of a country or area and thus does not involve any sampling procedures. For both, census and population-based surveys the data can only be as good as the quality of the questionnaire used. Both, for population-based surveys and census applies that not all members of a household are included. Commonly children aged 0-14 are excluded from the interviews.

[7] http://apps.who.int/ghodata/

Quality issues in census data

National census data are relevant for child mortality[8], fertility rates, birth rates and population data. National censuses should be carried out every 10 years. In area where civil registration is not covering at least 90% of deaths, mortality questions (child and adult mortality) should be included. It is standard to conduct a post enumeration survey in a selected subpopulation to assess the quality of census information. These data, together with the census data should be made available at least within 2 years of conducting the census. If mortality questions are included, additional reports on the quality of these estimations should be made available. Census reports include descriptive statistics by age and sex by the smallest administrative level and should be made publically available.

For additional models which are often added to censuses, the investigation of quality should cover the points as mentioned below for population-based surveys

Quality issues in population-based surveys

Target population

Most surveys only attempt to interview a certain part of the population. In DHS women and men of reproductive age are mostly included, while MICS (except a few new) only include women. Both, DHS and MICS only include 15-49 year old residents of households, thus the cohort of young adolescent aged 12-14 are not represented. Also, in some countries only women, who are married are included to answer question concerning reproduction. An assessment of the target population (the population for which estimates are required) included is thus in particular important for indicators concerning adolescents.

Key questions for assessment of the target population:

Is a target population clearly defined in the sampling methodology?

Is there a difference between the target population and the study population?

Does the study population cover the population needed for a construction of a respective indicator?

Sampling

The underlying concept of **representative population samples** is that participants should be selected at random (random sampling) and that everybody should have the same chance being selected (equal-probability selection methods). A simple random sample could be drawn if a list of all members of the population of interest is available. An example could be a school where the names of all pupils are available. The list of all pupils would then be the**sample frame** where a certain number of pupils defined by sample size calculation would be selected at random (e.g. using random numbers). However, most often a full sampling frame (list of all members of a population of interest) is not available, nor is

[8] In a few countries also maternal mortality data are derived from censuses

it economically feasible to visit many different places or villages.

That is why more complex sampling schemes are generally used, such as stratified, multi-stage and cluster sampling strategies. **Stratified sampling** is used, if the aim of a study is to have estimates for defined population groups, such as certain age groups. Most often **multi-stage sampling** is used. Here individuals are selected by, for example, first selecting districts, then villages and finally households (or districts, schools, classes). This reduces the need for full sampling frames of all population members. To select districts a list of all districts is needed, and then only for the districts selected a full list of all citizen in the villages becomes necessary. If the first units (in this example the districts) have roughly the same population size, simple random selection can be used. However if some are much larger in population size, the sampling technique needed is **probability proportional to size**. This means that for a district with twice the population, twice as many units may be selected. Cluster sampling methods are multi-stage sampling methods. Multi-stage sampling methods are less effective methods to obtain precise estimates compared to simple random sampling. It is expected, that within a cluster answers are more similar. For example, it is likely that within one village people share common characteristics including maybe difficult access to contraceptives. Thus the answers will be more similar to each other than a selection of people from different villages would have given. These 'sampling errors' can be estimated, and are expressed as a design effect. On overall the design effect rages from 1.5 to 2.5 for different variables and depending on the sampling strategy. The design effect is also important for sample size calculation. If multi-stage sampling is used the overall sample size has to be increased by the expected design effect.

Sample weight is used to adjust for the potential bias introduced when using multi-stage sampling. Thus sample weights outbalances that the sample taken was not fully at random and did not give all participants the same chance of being selected. This also becomes necessary if, after preparing the full list of population members within selected clusters, major differences in population size are found. This can be the case if major in or outmigration took place after the last census on which the enumeration areas are based[9].

Sample size and sample size calculation

A sufficient sample size is needed to provide the users with estimates of acceptable precision or to detect an effect of interest with a sufficient statistical power (see more in chapter 5.2.).

Sample size calculations need to be done for a range of key indicators of interest. For example, to detect a 10% increase in skilled birth attendance, uptake on antenatal care (4 times) around **500 women having had a live birth in defined period** would need to be interviewed. To detect an increase of 10 percentage point for contraceptive prevalence around **500 women of reproductive age** are to be interviewed. To detect a decline in child mortality with a reasonable precision more than **5000 women with a live birth** would need to be included.

[9] A good introduction is given in: Wassenich P. Data for Impact Evaluation. The World Bank. Doing Impact Evaluation No. 6. October 2007. *http://siteresources.worldbank.org/INTISPMA/Resources/383704-1146752240884/Doing_ie_series_06.pdf*

Sampling methods and sample size used in DHS and MICS

The **DHS** commonly use a multi-stage sampling procedure. A short description of the sampling strategy is part of the introduction. Sample errors are presented in more detail in the annex of the surveys. Moreover, for key indicators confidence intervals are given in the annex. The multi-stage sampling is in general based on enumeration areas, which are available from previous census. As a first stage, a large number of enumeration areas (clusters) are selected from these listed remuneration areas. A full list of all households in these clusters areas is then obtained and a defined number of households are selected using simple random methods, commonly around 20. In some DHS certain regions or urban areas are **oversampled** to get more precise estimates for these areas of particular interest. In response of oversampling, sample weights are given, so that the oversampling does not affect the overall estimates.

The sampling methods in **MICS** are comparable with the sampling methods in DHS. Also multi-stage sampling procedures are used and the sampling frame of the first-stage selection is also enumeration areas defined by previous census information. The sampling procedures are also shortly described in **MICS** country reports but quality tables are not found in reports.

DHS sampling methods and the overall sample size are commonly designed to give precise national and zonal estimates as well as values for urban and rural populations. In the annex B of DHS reports a table is commonly found listing standard errors and confidence limits for selected indicators. For example, in the Tanzanian DHS 2010, the total fertility rate was 5.4 with a confidence limit of 5.2 to 5.7. For urban areas the estimates were 3.7 (confidence limits 3.4-4.2) and for rural areas 6.1 (confidence limits 5.5-6.2). Estimation of the total fertility rate thus become imprecise if calculated for smaller areas, that is why for many indicators data only disaggregated for zonal but not regional level are given.

Table 1: Examples of confidence limits of relevant indicators from DHS 2010 Tanzania

Indicators estimates	National	Rural	Southern Zone
Contraceptive prevalence rate (modern)	0.27 (0.25-0.30)	0.25 (0.23-0.28)	0.40 (0.36-0.43)
Total fertility rate	5.4 (5.2-5.7)	6.1 (5.5-6.2)	4.4 (3.9-4.8)
Infant mortality rate (0-9 yrs preceding survey)	50.7* (44.1-57.3)	59.5 (53.3-65.7)	67.8 (50.2-85.4)
Skilled attendance	0.49 (0.45-0.53)	0.4 (0.37-0.44)	0.67 (0.58-0.75)
Condoms use last intercourse	0.27 (0.20-0.37)	0.26 (0.17-0.35)	0.29 (0.21-0.38)

(*data for birth 0-4 yrs preceding the survey, data for birth 0-9yrs preceding the survey not available)

Key questions for quality assessment of sampling procedure

Is the sampling method described?

Is multi-stage sampling used?

Is sample selected using probability-to-size sampling?

Is stratification used? If so, is the effect of stratification taking this into account in the analysis

(sample weights given)?

Is a sample size calculation presented?

Are confidence intervals presented?

Assessment of questionnaires and indicator construction

Assessment of questionnaires and single questions is important to access in particular **validity and reliability** of data.

Questions included are typically about 1) knowledge (what people know); 2) attitudes (what people say they want or think); 3) beliefs (what people say is true); 4) experiences (what has happened to people); 5) behaviors (what people do, have done, or intend to do); and finally 6) attributes (what people are).

To which extent questions are able to get valid (accurate, true) and reliable (answer does not change if respondent is asked again or another interviewer asks question) depends much on the wording and answer options. Questions need to be clear, not ambiguous, culturally sensitive, not threatening, not biased and answers should be mutually exclusive.

DHS and MICS survey both use well established questionnaire modules where the validity and reliability of questions have been assessed. Still, errors are known. For example, women report having undergone a Caesarean Section while other obstetric procedures, such as an episiotomy have been carried out[12]. During extensive data checking and cleaning procedures wrong answers can be found, for example by tabulating caesarean section with place of delivery. Both, DHS and MICS do extensive data quality checking and cleaning before the final reports are published.

A typical problem affecting child mortality estimations are heapings in reported age at death (clumping of respondents' ages on certain values). Commonly digit preferences are seen when examining information on age with preferences for values such as 5, 10, or 12. Age heapings can be a reason why estimations for infant mortality are surprisingly low compared with under-five mortality. Children dying for example at the age of 11 months are often misreported as 1 year and count thus as under-five-deaths but not infant death. To minimize errors, interviewers for DHS surveys are often instructed to record age at deaths in days for the 1st months and in months during the first 2 years.

Age heaping / clumping of respondents' age

Several studies have suggested that the fertility decline in several sub-African countries has stalled in the recent years. A re-examination of recent fertility trends in nine countries (Benin, Cameroon, Ghana, Nigeria, Rwanda, Tanzania, Uganda and Zambia) investigated in detail into the quality of data. When adjusting for misreporting such as age dis placement the study was only able to confirm a stall in Kenya´s fertility decline, and possibly in Benin, Rwanda and Zambia. But other stalls appeared to be overstated due to age displacement, omissions 25 and differences between DHS surveys in the late 1990s and those in early 2000s. This study highlights the importance of assessing data quality in particular when comparisons are made over time and questionnaire used are not identical. Age displacement / heaping were a major cause why a stall of fertility decline was suggested by previous studies [13].

Difficulties in getting reliable answers are also mainly seen for questions concerning sexual behaviors and sensitive issues such as alcohol use. Answers are often biased towards what the respondent thinks is expected and culturally acceptable (also described as information bias).

Also, recall periods need careful consideration. Mothers will remember events such as child deaths over a period of five years or even more, but data on age or months of deaths will be more blurred with longer recall periods (also described as recall bias). Much shorter recall periods are commonly used for coverage indicators, such as usage of ITN or tetanus prophylaxis during antenatal care.

Recall bias in out-of pocket expenditure assessments

Recall bias has been identified a major problem in the measurement of out-of-pocket expenses. An investigation based on World Health Survey which included two questions on the cost of hospitalization, one with a 4-week and the other with a 11-month recall period revealed that the estimations derived from a shorter recall period are commonly more than two or three times higher. The highest ratios of around nine-fold higher expenditure stated with 4-week recall period were observed in Uruguay, The United Arab Emirates and Swaziland. In addition differences are seen whether ‘overall’ health expenditure or expenditure per different categories such as outpatient, traditional medicine, dentist, medicine and others was assessed. The single-item measurement yields a significant lowerestimate [14].

Recall periods and categorizations assessed vary between survey types. In the World Health Surveys, a 4-week for outpatient and 11-month recall period for in-patient care was used. The World Banks Living Standard Measurement Surveys use different recall periods, generally up to 12 months. Here also categories used differ between countries. Thus comparisons of estimation of out-of-pocket expenditure derived from different types of surveys are meaningless.

Household surveys should either use pretested questions or questionnaires (as DHS and MICS do) or need to involve in a series of questionnaire development procedure. This includes pretesting and piloting. Pretesting describes a first check with a few respondents whether questions are clear, answer options sufficient, the flow of the questions is appropriate, recall periods are sensible, length of questionnaire is manageable and others. A pilot refers to a test of the questionnaire with a larger number of respondents and a test of field procedures.

Key questions in questionnaire design

Has a standardised, pre-validated questionnaire instrument been used?

If no standardised, pre-validated questionnaire was used:

Was the questionnaire designed together with relevant stakeholders and resource persons?

Was the questionnaire pre-tested (few interviewees)?

Was a pilot done (selected sites, several interviewees)?

Has the questionnaire layout been checked (flow of questions, instructions for interviewer,

appropriate coding, uniform codes for missing values and non-response?

Has the questionnaire been administered in another language then originally designed?

If so: Has the questionnaire been translated and back-translated?

Completeness of data

Completeness is another important quality criterion. Completeness is important to evaluate 'generalizability' of results and to exclude major selection bias (selection bias is introduced when the group of people actually interviewed is different from the group selected for interview). Both, the overall response rate and the response rate to certain questions are of importance.
Data or participants can be missing at random or non-random. If data or participants are missing at random (or in other words just by chance) it is assumed that no major bias is introduced. However, it is rare that data are really missing at random. Mostly it must be assumed, that certain characteristic of the respondents have led to non-participation or refusal to participate. This can lead to bias as the group of respondents and non-respondent might differ, e.g. they could belong to different socio-economic groups. A sensible way to investigate into potential bias introduced due to non-response is to use all available information and household characteristics and to compare respondents and non-respondents. If comparison suggests that respondents and non-respondents are similar (similar distribution of education level, wealth, urban rural, etc.) a major selection bias can be excluded.

DHS and MICS report relatively high response rates in most countries. The DHS Tanzania 2010 reported a response rate of 96% for eligible women and 91% for eligible men. This high response rate is achieved by re

peated visits to the household. Still response rates for men are slightly lower which reflects the more frequent and longer absences for men. DHS reports commonly present a table in the introduction section where response rates are tabulated for urban / rural distribution. Response rates found for DHS and MICS are rated as sufficient and no major bias has to be assumed.

Non-response is a particular issue for taking blood samples for HIV-testing in AIDS indicator surveys. In the latest AIS from Tanzania, 6% of women and 8% of men interviewed refused to provide blood for HIV testing. Although the percentage of refusal of 6-8% is in the same range as the overall non-response, caution is here required: As these women and men participated in the principal survey socio-demographic characteristic are known. The final report of the AIS TZA 2007/08 includes a table giving background characteristics of people being tested and people who refused. The results suggest that a 'non-random' error was introduced by the refusal. More women and men in the highest socio-economic group did refuse to take part in the anonymous testing. HIV prevalence is little higher in the group of people in the highest wealth quintile, suggesting that the overall population result is slightly biased to an underestimation because not all selected participants gave blood samples.

Thus, results from HIV testing have to be viewed with some caution, but as for this example, the total number of people who refused was relatively small, 'true' results might not be very different. In epidemiological studies, sensitivity analyses are sometimes carried out. Sensitivity analysis test assumptions such as what would be the overall prevalence if all respondents who declined were positive or negative or have the same prevalence as their background characteristics suggest.

Key questions in completeness

Is information given about how many people were intended to be included and how many were in fact interviewed?

Is the overall response rate over at least 80%, better 90%?

Are background characteristics available for non-respondents / people who refused to participate?

If so is the distribution of background characteristics similar in respondent and non-respondents?

4.3. Quality assessment of routine health service data and administrative data

In most low income countries public health surveillance or health services data (mostly referred to as health management information system (HMIS) were introduced more than 15 years back. Many of these routine health service data systems evolved from epidemiological surveillance systems which were mainly focused on disease control, thus assessing rather the occurrence of a few selected diseases, such as cases of poliomyelitis, neonatal tetanus or cholera for outbreak monitoring. At that time surveillance systems were allowed to be incomplete as rather a steep raise in cases would alert action, whereas the actual count was not primarily important.

With continuing decentralization and health reform processes, a further objective of routine health information system was added, being the provision of local data for health planning, which raised expectations in completeness and overall quality. But the need for high quality, complete and timely health information from routine service provision rose rather with the needs of global initiatives and global partnerships, and the introduction of pay-for-performance systems.

It is important to bear these recent developments and the historical perspective in mind as many countries are only slowly revising their routine HMIS to adapt to new demands. In Tanzania, for example, the actual version of the health information system (HMIS or MTUHA in Swahili) dates back to the 1990s, and only minor adaptation were done throughout the past 10 years. A new version is only expected to be released soon.

HMIS have advantages and disadvantages compared with surveys or epidemiological studies. **Advantages include**

- Documentation is an ongoing activity and data can be provided timely (and not only every 4-5 years as DHS/MICS)
- Information is available for single health facilities, districts and regions/zones
- Early problem identification and rapid action
- Documentation is part of the routine system, thus extra costs are judged minor

Disadvantages include

- Amount of details which are collected is restricted
- Quality of data is limited
- Completeness is often low

HMIS or health service data are clearly complementary to population-based surveys, or other resources. There major advantage is that HMIS provides data for the district level and in regular intervals and can thus assist greatly in monitoring health programmes at district level.

Main indicators from HMIS

Many global initiatives are relying heavily on information collected as part of HMIS or through separate forms typically introduced for vaccination, HIV/AIDS and TB programmes. Many UNGASS-Indicators are based on service output information such as for example of two indicators which are also used in GDC for monitoring progress in Control of HIV/AIDS and other sexually transmitted diseases):

- Percentage of adults and children with advanced HIV infection receiving antiretroviral therapy
- Percentage of HIV-positive pregnant women who receive antiretroviral medicines to reduce the risk of mother-to-child transmission

Also monitoring of progress in tuberculosis depends on documentation of TB cases at the service delivery level. Examples of indicators used also for GDP are:

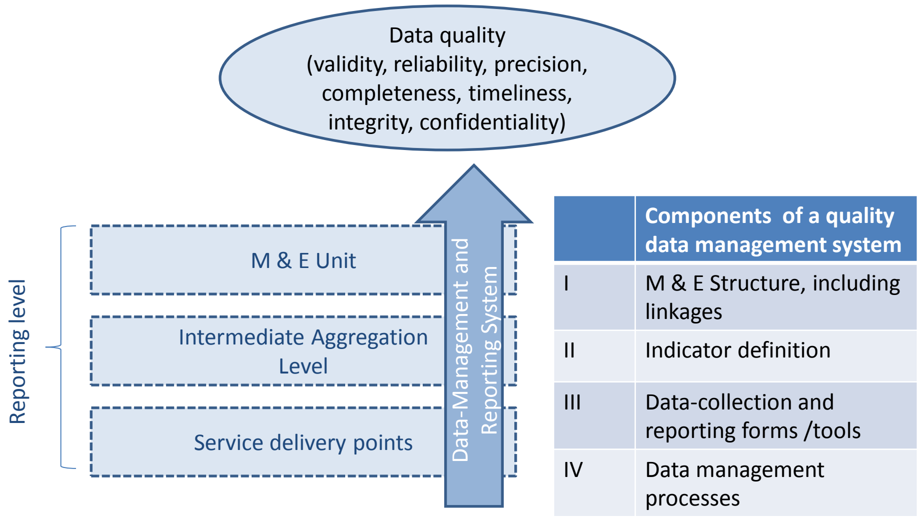
- Incidence of tuberculosis (per 100 000 population per year)
- Tuberculosis treatment success under direct observed treatment (DOTS) (%)

Also many other indicators relevant to other thematic clusters might be available from routine HMIS. Investigation into available data and indicator definition beforehand might greatly reduce the need for special surveys for results assessment in GDP.

Framework to assess national routine HMIS

With the increasing importance of routine health service data for global initiatives work has been done to assess the overall quality of data derived from these HMIS system or parallel recording at health facility. USAID has supported an comprehensive assessment tool [15] and also the Stop TB department has drafted a detailed manual on how to audit the quality of key TB indicators [16].

The data quality assessment tool published by USAID provides a useful overall framework and aims at 1) verifying the quality of data 2) assessing the systems that produces the data, and 3) develop action plans to improve both.


Data quality
(validity, reliability, precision, completeness, timeliness, integrity, confidentiality)
Reporting level
M & E Unit
Intermediate Aggregation Level
Service delivery points
Data-Management and Reporting System

	Components of a quality data management system
I	M & E Structure, including linkages
II	Indicator definition
III	Data-collection and reporting forms /tools
IV	Data management processes

Figure 2: Framework to assess routine health data, with minor adaptations from [15].

Assessment of the overall M&E structure, functions and capabilities and linkages between systems

As routine health data systems are ongoing activities they need established structures, with defined functions and adequate capabilities and time resources. A quality assessment of these structures need to include the assessment of organizational arrangement, managerial functions and staffing (sufficient, adequately resourced), protocols, national M&E plans, and most important data flow charts. This is relevant at central, but also intermediate and lower level.

It is likely, that several M&E units need to be assessed. HIV/AIDS or TB control programmes, multi-sectoral HIV programmes, vaccination and family planning/reproductive health units might have independent M&E units, dependant on level of integration of specific disease control efforts in the countries

Particular attention might need to be given to mechanisms to include the private (for-profit and non-profit) sector in the overall monitoring arrangement. To which extent they are included might vary between systems.

The existence of data flow charts and clear documentation of aggregation levels and responsibilities is a quality criterion. Data flow chart might indicate where double reporting could be a problem. Double reporting is described particularly for vaccination services where vaccinations done during child health days might be directly reported to the central level but also be documented by health workers as part of their monthly reporting. Data flow chart can also assist to assess how well systems are linked to each other and exchange of information is established at all levels of aggregation and use of data.

Health service data /HMIS in Tanzania

In Tanzania several parallel recording systems at health facility level are established. All providers in the health facility fill the common patient registers and summary reports being part of the national HMIS (called MTUHA in Tanzania) which includes reporting requirements for the immunization as well for the reproductive health programme. However, no distinct health provider at the first-line facilities is commonly charged with the responsibilities to prepare the reports and held accountable. In addition separate systems for recording were established for TB, STI/HIV/AIDS and PMTCT. These reports are based on separate patients and laboratory registers. For the disease specific reporting clear responsibilities have been established at health facility level but also at district level, which is not the case for the overall HMIS system where data compilation or follow up of missing data is not the explicit responsibility of any member of the district health management team. The HMIS data despite deficiency enjoy a high credibility among district health stakeholders whereas NGOs or donors often feel that they provide data of to low quality.

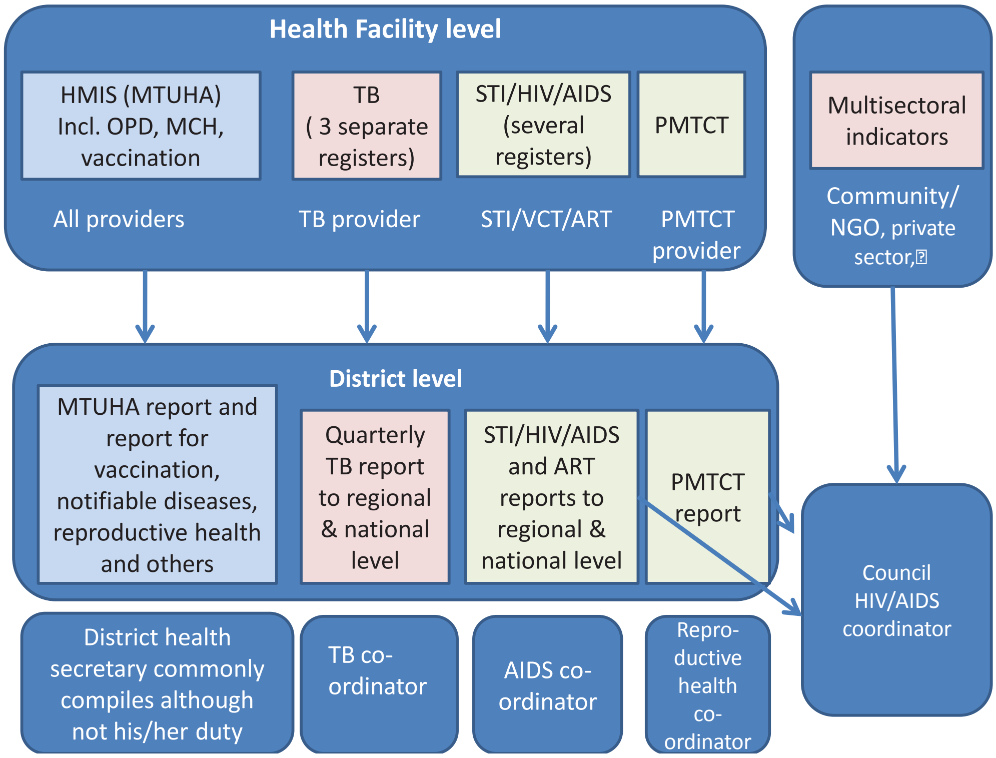

Health Facility level
HMIS (MTUHA) Incl. OPD, MCH, vaccination
TB (3 separate registers)
STI/HIV/AIDS (several registers)
PMTCT
All providers
TB provider
STI/VCT/ART
PMTCT provider
Multisectoral indicators
Community/ NGO, private sector,
District level
MTUHA report and report for vaccination, notifiable diseases, reproductive health and others
Quarterly TB report to regional & national level
STI/HIV/AIDS and ART reports to regional & national level
PMTCT report
Council HIV/AIDS coordinator
District health secretary commonly compiles although not his/her duty
TB co-ordinator
AIDS co-ordinator
Repro-ductive health co-ordinator

Fig 3 : Parallel reporting and responsibilities of health facility data

Key questions in assessing structure of health service data collection /HMIS

Is one common recording system /HMIS used throughout the country?

Does the system include private-for-profit and non-profit providers? If several systems are used:

Are mechanisms established to link separate systems?

Are mechanisms established to reduce double recording?

If several recording systems are established in a country, quality assessment would need to be carried out for each of these systems.

Are responsibilities for recording clearly outlined at first-line health facilities?

Are responsibilities for summarizing / reporting clearly outlined at district /first aggregation level?

Are responsibilities for summarizing / reporting clearly outlined at 2nd and 3rd aggregation level?

Is sufficient time allocated for recoding and aggregation of all health providers / managers involved?

National level

Has the national M&E unit sufficient capacity in terms of staffing and qualification of staff?

Are data flow charts available?

Are clear guidelines available concerning collection, aggregation, responsibilities and time line?

Is data quality audits part of the tasks of the national M&E unit?

Assessment of indicator definitions / case definitions

Another quality criterion for routine health data systems is that case definitions and additional explanations for recording are available to the people in charge of recording. Assessments in health facilities can help to verify whether definitions are clear and locally adequate.

For the purpose of use of HMIS information for results assessment GDP it is also important to verify whether the indicator definitions are the same or at least similar to what is intended to measure and used in frameworks.

National definition of indicators might differ within information systems, particularly between routine health reporting systems and population-based surveys. Examples of deviant definitions are for example the reporting of 'health facility births' whereas the indicator commonly discussed is 'percentage of birth with skilled birth attendant'. In countries, where all staff categories are judged as skilled and health providers neither attend birth at home nor in private maternities, information on 'health facility birth' can be acceptable for the construction of the indicator 'skilled attendance'. But depending on the context and the number of deliveries done by community midwifes, the uses of 'health facility birth' might under- or overestimate the percentage of women delivering with a skilled attendant.

Much effort in clear definition of indicators has been put into the monitoring of TB[10] and notifiable diseases as well as into monitoring of vaccination coverage. Drastic drops or increases in prevalence when monitoring over time should always raise alert to look for changes in case definitions or availability of new technology for case detection. The recent rapid drop in malaria cases where malaria rapid tests are becoming increasingly available underlines the need to carefully assess case definitions and actual practice for diagnostics.

Nominator/denominator issues in coverage indicators

A major problem when using coverage indicators from HMIS are nominator/denominator problems. The nominator for coverage indicators is based on events that occurred in health facilities. But the denominator is based on the overall population (or expected birth) in the target area of a health facility or the district. The overall population data are commonly based on projections from census data. Thus nominator and denominator come from different populations.

Several problems are relevant. First, defined target areas of health facilities might not overlap with health care seeking behavior of the population because people might use neighboring or higher level facilities. Second population projections can be greatly wrong, in particular when the census data are more than 5 years old, or migration is common. Thirdly, and this applies for data using expected birth, the national birth rates (this is what is commonly used) might differ from local birth rates. Still, many indicators might be helpful for results assessment and data should not be disregarded. But it is important to note that coverage indicators obtained from HMIS might differ greatly from what is observed in population-based surveys. Indicators cannot be used interchangeable. And the respective source of data should be given when presenting data.

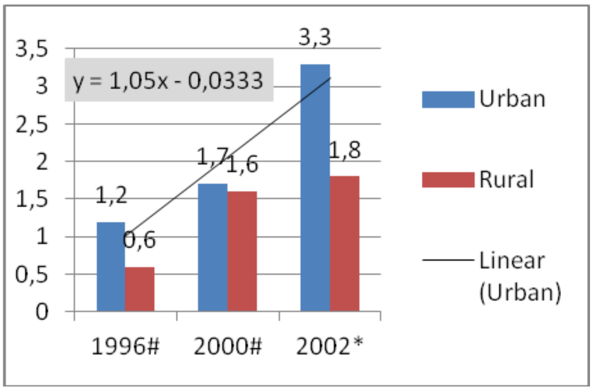


Fig. 4

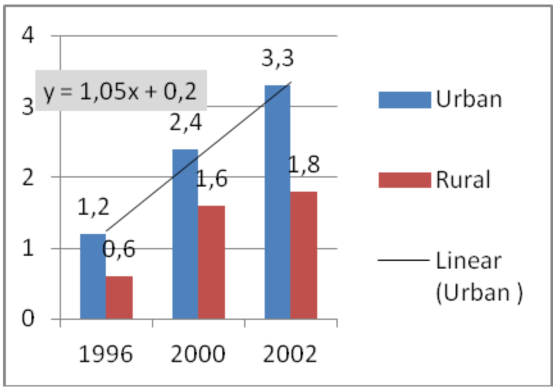


Fig. 5

The figures above present an example of difficulties to estimate coverage levels between censuses. These graphs present data from levels of major obstetric interventions (primarily Cesarean Section) per estimated births which is an estimate of access to emergency obstetric care from a study done in Mtwara region, Southern Tanzania by GIZ [17]. Censuses were carried out 1988 and 2002 in Tanzania. Figure 4 is based on expected birth estimates based on the projected population estimates from census 1988 for the years 1996 and 2000 whereas in figure 5 the levels where re-calculated based on the results from the 2002 census. Both figures show clearly that the coverage levels

are improving suggesting improved access to life-saving operations, but the level in the year 2000 for the urban population might have been too low because projections from the 1996 census were not able to account for the change in population pattern.

Still the example also indicates that despite the nominator/denominator problem information from routine recording can be useful to observe overall trends, even if absolute levels are not likely to be exact. But more advanced statistical analysis should not be done because of likely error in the estimations.

Data management process

The assessment of the data management process aims in particular to look for completeness. Cross-checking between different registers, laboratory or patients notes and recalculation of summary sheets are common activities to assess completeness and also reliability of the data management system. Assessments should also verify whether reporting forms and registers are in fact available where they are needed, and whether they are user-friendly and support documentation.

Assessment of completeness at the 1st aggregation level (district level) includes a check whether all health facilities do report, and reporting is timely. It also includes a reassessment whether the aggregation at 1st level is correctly calculated and includes all necessary information. The aggregation procedures should be of best use with computerized systems which facilitates aggregation at higher level, but the reality is often that printed tables are produced, and these figures need to be entered and aggregated again at the next aggregation level which inherently risks to lead to calculation errors. Assessment at all levels should also encompass the issue of double reporting. Systems of good quality should have a regular check for completeness established as a continuous activity [16].

Some assessment tools guide the users to exactly quality the completeness of the data, which can help a lot because uncertainty intervals could be calculated when used for result measurement.

Key question in assessing the management and recording process within HMIS

Are registers and reporting forms continuously available?

Are summary sheets user-friendly and continuously available where they are needed?

Do counts from different registers and summary sheets tally at first-line health facilities?

Is aggregation at 1st and 2nd aggregation level correct (do counts tally)?

Are PC used for aggregation to produce printed tables?

Are data entry systems used which assist for producing summary sheets at 2nd or 3rd aggregation level?

Use of coverage data from HMIS data and population based data for monitoring

n most countries population based coverage data such as for births attendant by a skilled provider, antenatal care attendance or vaccination coverage are only available every 5 years when new DHS data are released. However,

[10] See more: *http://apps.who.int/tb/surveillanceworkshop/*

these intervals rarely coincide with the reporting requirement of project and programmes. To fill these gaps often estimates from HMIS and DHS are used interchangeably, irrespective of differences in definitions and quality issues in each of these methods. HMIS data often over or underreport coverage, depending on how complete the compilation of records from health facilities and how important the denominator is.

Information on skilled attendance from Southern Tanzania from two sources: HMIS and DHS

The following two graphs give information on birth in health facilities and by a skilled provider from two different sources: HMIS made available from Zonal Reproductive Health Report [18] and the DHS 2004/05 and 2010. Estimates given in figure 6 use the definition of 'births in health facilities'. In Southern Tanzania, almost all birth in health facilities are by a skilled provider and home birth by professionals are little encouraged.

Births in Health Facility (HF) Lindi and Mtwara region combined (Source: HMIS)

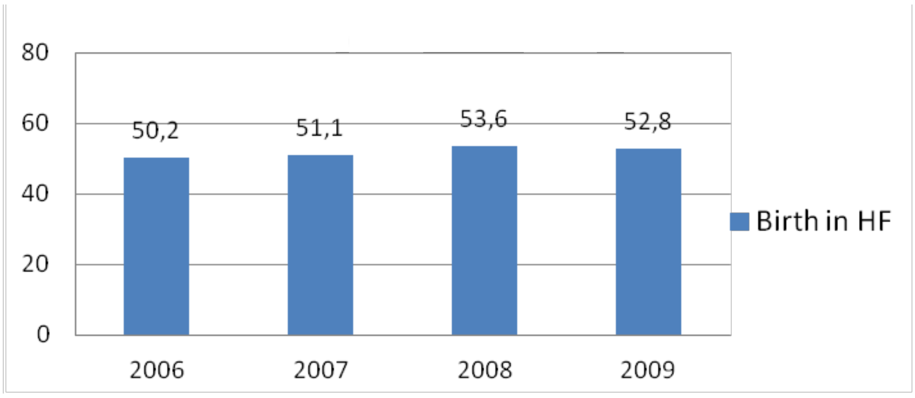


Figure 6: Birth in health facilities based on HMIS data

Births in Health Facilties (HF) and by skilled provider in Lindi and Mtwara combined (Source: DHS 2004/05 & DHS 2010

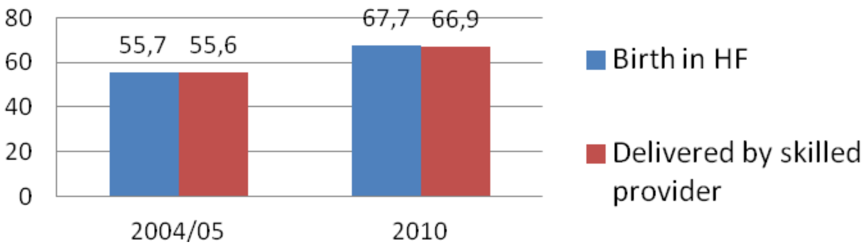


Figure 7: Birth in health facilities based on DHS data

The estimates from the DHS (Figure 7) show a clear increase in both, births in health facilities and births attended by a skilled provider. If the information would have been used interchangeably, thus DHS estimations for the year 2005 and HMIS- estimations for 2008, the increase would have been missed. Both information sources are helpful but estimation should not be used within one table or graph or to show trends

Quality assessment of health accounts

Data available from national health accounts are derived from several sources, government accounts, donor information on funding and population-based surveys investigating into private expenses for health. Quality assessment of health account thus has to investigate into the primary sources of this financial information.
In principle, assessment of information of government health spending might follow similar steps as presented as part of the assessment of health service records/HMIS thus investigation into structure, management and definitions/categorization. In addition, the overall legal financial framework in a country will be major determinants of the reliability of the data and needs to be included in the assessment. Also audit report will give important information. A paper prepared by the IMF gives here[11] more guidance.

Information on household expenses on health is derived from population based surveys. Here an assessment of the questions used to obtain data on health expenses as questions and recall periods differ between surveys as discussed before.

4.4 Assessment of vital registration systems

Vital registration (also called civil registration, describing a system of continuous registration of all birth and deaths in a country) is the gold standard to obtain information on causes of deaths. Deaths certification should be based on the "International Classification of Diseases and Related Health Problems" [19]. The 10th revision was issued in 1994, but many countries continue to use the 9th revision. Main quality criteria for assessing vital registration systems are the completeness, coverage and quality of information of causes of deaths. Completeness refers to the extent to which deaths and causes of deaths are recorded for the whole population - for young and old age, urban and rural population. Comparison of death distribution between census data and vital registration systems can assist to quantify and correct for incompleteness [20]. Coverage refers to the extent, all geographical areas are included in the registration systems, and that guest workers or refugees are not excluded. The quality of information on causes of deaths is commonly defined by the proportion of deaths coded to 'unknown' and 'ill-defined causes' [21].
In 2005, Mathers et al. [21] assessed the deaths registration systems of all 115 countries in the world where birth and deaths are routinely recorded. Essentially only the deaths registration from high income countries, but also Mexico and Venezuela were rated as high quality where completeness of deaths registration was estimated to be above 90% and ill-defined causes of deaths were only used for less than 10% of deaths. For most countries deaths registration systems, including the German registration system, as well as that of many other European countries were rated as medium quality particularly as more than 10% of deaths were coded as 'ill-defined' causes. Ill-defined causes of death included here the use of 'symptoms' or cardiovascular disease categories lacking diagnostic meaning such as 'cardiac arrest' or 'heart failure'.

Several investigations into the reliability of estimates of causes of deaths has been done for maternal mortality data derived from vital registration systems and major underreporting due to misclassification has been found. Two studies in the

[11] IMF. The fourth review of the fund's data standards initiative. Supplement to the data quality assessment framework. Pre pared by the Statistics Department, 2001

Netherlands calculated the level of underreporting at 26% for the period 1983-1992 and 33% for the period 1993-2005 [22, 23]. In Finland 8 out of 11 (72%), in France 30 out of 83 (36%) and in the two American states 31 out of 58 (53%) maternal deaths were not reported in the official statistics, but only found through data linkage of birth and deaths registers [24]. Also in Austria, 38% of maternal deaths were not reported as maternal [25]. Misclassification is a problem in particular for eclampsia and indirect maternal deaths such as deaths due to underlying causes aggravated during pregnancy. The most recent global estimates of maternal mortality have adjusted the maternal mortality data from countries where estimates were based on vital registration to account for misclassification and underreporting in vital registration systems [26, 27].

4.5 Published data /literature review

Much country data on impact indicators have been published in 'The Lancet' journal over the past years. These include the latest estimation of maternal mortality from 181 countries [26], estimation of stillbirth rates [28], estimations on child and infant mortality [29] including estimations on cause of deaths and estimations of neonatal and post-neonatal mortality [30].

Moreover, the Lancet published several thematic series during the past years on child health, reproductive health, maternal heath and other important health issues and thus became an important data source. Some of these articles and series can be retrieved free of charge[12].

It is important to note that the estimations published in peer reviewed articles - although often based on available secondary data such as DHS, MICS or vital registration – differ much from original estimates as they have been subject to extensive adjustments and modeling exercises to account for measurements problems such as underreporting or misclassification, sample errors and other data errors. The estimations on infant and child mortality have been relatively consistent, but not so the estimations on maternal mortality. The partly very different country estimation of levels for maternal mortality published in 2010 from the Institute of Health Metrics [26] and from the WHO, UNICEF, UNFPA and World Bank [27] underline the need for increased transparency in modeling strategy and better collaboration within scientific and country experts. Sadly, none of the two available estimations can be recommended for country monitoring as the estimates differ greatly at country level [31].

Many publications are today freely available. A search in PubMed can identify a wealth of data[13] . The search strategy can use topics and geographical area; however this often leads to a large amount of articles. Another possibility is to use author's names as groups of researcher within a certain thematic area are mostly known in the countries. By this a large amount of additional information can be collected to describe progress or deficiencies in more detail.

[12] See www.scholar.google.com

[13] *http://www.ncbi.nlm.nih.gov/pubmed/*

5. Primary data collection for results measurement

Primary data collection should only take place when there are no adequate secondary data to monitor progress. Primary data collection might also become necessary if the effect of an intervention should be measured rigorously thus impact evaluation with a comparison (counterfactual) is needed. Here secondary data sources might not provide all necessary information, or the data quality from available data sources might be too low or the available data sources do not cover the right periods before and after the intervention and the intervention and comparison areas.   This chapter only aims at giving a very rough overview of issues in preparation, methods, analysis and documentation, more detailed information are available from other sources[14] and particularly textbooks. They are described in more detail in the annex.

Moreover as sampling and issues in questionnaire design have largely been covered in chapter 4 and information here is rather complementary.

5.1. Preparation of primary data collection

If primary data collection is needed, enough time and care needs to be invested in the planning and preparation phase. Data generation should not be done in isolation; instead it should include relevant stakeholders and available expertise. Consultation of relevant literature and in country experience is of utmost importance. Sound knowledge of available published and unpublished literature is needed to draft a proposal and study protocol for ethical clearance which in general has to include relevance, major objectives, methods and tools used.

The overall framework for the analysis determines which data are needed typically. In general 1) a primary outcome variable (also called dependant variables often health outcome or impact indicators) need to be chosen. The primary outcome variable relates to the main subject of the study. In addition, secondary outcomes variables (other relevant output, outcome and impact indicators), and 2) independent variables (also called explanatory variables, typically socio-demographic indicators such as age, wealth-status) and variables of confounders need to be defined.

The level of observation needs to be determined, whether individual, household or community level. If interventions are directed to communities rather than individuals, then the evaluation might consider rather community than individual level variables.

If the variables are defined, the most appropriate approach – quantitative and/or qualitative - and the best methods for obtaining them need to be decided upon.

Also, the best timing of baseline and follow up assessment needs consideration. Although the literature stresses, that baseline surveys need to be done before the onset of an intervention, some variables inherently measure periods before the survey is carried out. These include infant and child mortality data and coverage indicator which are based on 'birth histories' covering a period of 2 to 5 years prior a survey. The period of 2 or 5 years is important to have a sufficient case number included in a sample. In contrast, indicators of knowledge (such as knowledge of HIV prevention measures) and

[14] The following sources give more detailed guidance: Wassenich P. Data for Impact Evaluation. The World Bank. Doing Impact Evaluation No. 6. October 2007. http://siteresources.worldbank.org/INTISPMA/Resources/383704-1146752240884/Doing_ie_series_06.pdf & Dicklberger et al, Baselineerhebung. Ein Leitfaden zur Planung, Durchführung, Auswertung und Nutzung der Ergebnisse, GTZ 2010
Slides prepared by The World Bank and Sistema Intergrales for Cuernavaca, Mexico available http://www.impactevaluation2011.org/forum/wp-content/uploads/2011/06/C-8-file-1.pdf International Household Survey Network. Survey Quality Assessment Framework SQAF. Prepared by the Statistical Service Centre of the University of Reading. http://www.surveynetwork.org/home/index.php?

attitude refer to the time the survey is carried out. Thus for mortality data and coverage indicators it is important that the evaluation is not done immediately after the intervention was implemented but with an appropriate time lag.

For all studies a team of interviewers and researchers will need to be employed. Survey teams are typically composed of interviewers, a supervisor (or field manager), a data specialist, data entry clerks and an overall survey manager. All members of the team need training or at least a thorough introduction to this specific study procedure, formal training or on-the-job training. Training material needs to be developed.

For larger surveys new technologies such as Personal Digital Assistance (PDA) provide the opportunity to enter data directly instead of first using paper based instruments. Using this new technologies helps to ensure quality and data integrity and makes information more timely available [32].

Budgets need to cover all costs; salaries and allowances, training, equipment, transport, stationeries and cost for reports/publications need to be included. Funding of all costs needs to be ensured before any data collection is started.
In sum, enough time and care should be devoted to the planning phase for preparation, sharing and discussing with all relevant stakeholders, budgeting, planning, purchase of necessary items and training of interviewers. Ethical clearance is often taking several months, hence delaying the start of any research activity.

Key questions in the preparation phase

Are relevant stakeholders and experts participating in the design of the study and which role do they have?

Have relevant data sources / the literature been consulted?

Is ethical clearance obtained?

Are budgets comprehensive and reasonable?

Are funds sufficient?

Is a survey team available with sufficient capacity?

Are clear job description / task definitions drafted?

Is training material available?

Is a training planned /done?

Are all necessary items purchased?

Are quality assurance measures planned?

5.2. Sampling and instruments for data collection

The first and foremost important question in sampling is for which target group relevant information or estimates need to be obtained. If adolescents are the main target group, then the omission of the age group 13-14 as in most surveys of reproductive age presents a serious limitation.

Sampling and sample size

Sampling methods are relevant to quantitative and qualitative methods respectively. Quantitative data collection mostly aims at obtaining a representative estimate for a defined population group. Here sampling methods using equal probability sampling methods (simple random sampling and multi-stage sampling) are needed (see 3. 2). Other sampling methods might be used if the primary aim is not to get a representative estimate but rather qualitative information. Less systematic sampling methods are also acceptable for communities which are hard to reach such as drug users or sex workers.

Purposeful sampling methods or other methods not strictly including participants at random are appropriate for research which does not aim to give representative results but rather investigate into qualitative aspects. These are much used to investigate into the quality of care provided in health facilities. Purposeful sampling can save costs and time, and might in some settings be more ethically acceptable. However, important for the final discussion of results is, that these finding cannot be generalised to a larger population or other health facilities. The information obtained gives strictly speaking only valid estimates for the respective group / facility which were included.

Chain-referral sampling such as snowball or respondents-drive sampling is one purposeful method where participants are asked to recruit their peers. These methods have been much employed in surveys directed at hard to reach groups such as injecting drug users [33]. These methods are able to sample participants from groups where probability sampling methods are impractical because of the lack of an adequate sampling frame. Although selection bias has to be assumed because not all segments of the target group might be reached, in general these sampling strategies are rated as effective [34].

Sample size estimations need to be done for the main (primary) outcome and secondary outcome variables. To calculate the sample size the right formula needs to be chosen. Standard formulas differ for different study designs and whether the aim is to compare means, rates, and proportions. Some decision needs to be taken before this calculation. A value for the ‘precision’ (the desired width of the confidence interval (CI)) needs to be agreed. Commonly a 95% CI is accepted. The ‘power’ of the study defines the probability of detecting a significant result. Typically a 80% probability is chosen. For comparison of means, rates and proportions the likely values need to be known as well as the expected difference between the values. Some statistical programmes have an option for sample size calculations, but it is advisable to contact a statistician in particular if more complicated study designs are planned.

Important is that the sample size calculated need to be adjusted for the sampling strategy (design factor). A rule of thumb is that if cluster sampling is applied the overall sample size need to be double the calculated size. Moreover the final sample size should include a contingency if some selected participants cannot be reached.

Sample size calculations are based on the effect size of interest (e.g. did skilled attendance increase by 10 percent points, or from 40% to 50%), standard deviation (statistical measure for which common values are used), confidence level (commonly 95% confidence levels are used which mean that one can be 95% sure that the obtained estimates lies within two standard deviations of the ‘true’ population value), power level (commonly set at 80% meaning that in 80% of studies a significant effect is not missed), and finally the design effect (commonly around two in multi-stage sampling).

Key questions in target population, sampling strategy and sample size

Is the target population clearly defined?

Is the sampling method chosen in response to the objective of the study?

Is the sampling strategy clearly defined?

If multi-stage sampling is used: What is the basis for the first selection process (sampling unit 1st stage of sampling)?

Is proportional to size sampling done?

What is the basis for all following sampling stages (sampling unit, selection mode)?

Are full sampling frames available?

Was a sample size calculation done for primary and secondary outcomes?

Was the sample size adjusted for sampling strategy and 'lost' observations?

Instruments for data collection

Instruments used in quantitative research can include questionnaires (administered such as household interviews, self-administered, structured, semi-structured,..), diaries, observation checklists, and others. Administered questionnaires are used for most population-based surveys in low and middle income countries. They have the advantage to ensure a high overall response rate and low non-response to single question. Self-administered questionnaires are often used for sensitive questions to minimise the social desirability bias. Social desirability bias describes the tendency of respondents to reply in a manner that will be viewed favorably by others. This will generally take the form of over-reporting for perceived good behaviour or underreporting bad behaviour.

Questionnaire design and layout is a challenging task. Questionnaire design can heavily influence the data quality. Quality in questionnaire development is best ensured by using pretested and validated questionnaires or questions. If this is not possible it is advisable to draft questionnaires together with a group of people, including data managers and the people who will do the analysis. Questions should be brief, objective, simple and specific. The sequence and the flow of questions can have a great impact how they are answered and need equally attention[15].

Questionnaires need to be piloted and pretested, and if questionnaires are administered in another language than designed, they need to be translated and back-translated to ensure that questions are formulated in a way that the responses give valid information (see 4.2).

Key questions in instrument choice and questionnaire design

Is the instrument chosen the best way to obtain the needed information (validity)?

Has the rationale to use the chosen instrument been documented?

[15] See more detailed advice and detailed information in Wassenich P. Data for Impact Evaluation. The World Bank. Doing Im pact Evaluation No. 6. October 2007 &
Emily White, Bruce K Armstrong, Rodolfo Saracci. Principles of Exposure Measurement in Epidemiology: Collecting, Evaluating and Improving Measures of Disease Risk Factors, 2008.

Will a standardized, validated instrument/questionnaire instrument been used?

If not: Is the instrument/questionnaire designed together with relevant stakeholders and resource persons, including data managers and persons involved in the analysis?

Are questions brief, objective, simple and specific?

Are answer options exhaustive and mutually exclusive?

Was the questionnaire pre-tested (how many interviewees)?

Was a pilot done (selected sites, several interviewees)?

Has the questionnaire layout been checked (flow of questions, instructions for interviewer, appropriate coding, uniform codes for missing values and non-response)?

Will the questionnaire be administered in another language than originally designed? If so: Has the questionnaire been translated and back-translated (i.e. double-checked)?

5.3. Field work organisation, data management, data analysis

Field organisation and data management are important factors to ensure completeness and also have an effect on the overall quality of data. Key elements of quality assurance are 1) the composition of and the qualification within the team and 2) continuous quality assurance mechanisms including the availability of field instruction manuals.

Both, the previous experience in data collection and the training provided as part of the preparation for data collection is important in ensuring good capacity to carry out the planned data collection. Training should 1) give enough background information on the overall goal and objectives of the study 2) provide a thorough guide through the data collection instruments and 3) might include, some medical training, to enable, for example, identification of medical instruments.

Continuous quality assurance mechanisms also include the layout and design of the questionnaire. Clear skip patterns and predefined ranges of values increase the quality of data. For example, the interviewer needs to be clearly told that if the answer to a question was 'no', he/she should continue with question XY. Ranges of value should be mentioned e.g. 1-10. Continuous quality assurance is much enhanced by using PDAs where skip patterns are pre-programmed (questionnaire continues automatically with next questionnaire topic if the questions are not applicable because of a 'no' answer. Moreover, ranges of values can be pre-coded, and consistency and logical checks are done while interviewing. Regular daily summary sheets including number of interviews/observations done, calculation of means of key indicators and comparison of distribution of values among interviewers can greatly assist to ensure quality. Commonly the field manager or field supervisor takes over the main responsibility that data collection is done according to plans, summary sheets are available and daily checks are done.

Field instruction manuals provide guidelines and instructions for the interviewers. The content should include 1) why the survey is done, when, how the information will be used, and who the main investigator is; 2) sampling strategy and procedures; 3) ethical principles, consent and ensuring of confidentiality; 4) the role of the interviewer, introduction, advise to ensure high response rates; 5) instructions how to go about the questionnaires including a description of formats, skip patterns, and coding; 6) probing and the use of feedback.

Data should be stored at different places to prevent loss of data. Daily back-ups on PC/storage medium or preparation of photocopies are essential. Moreover, back-ups and photocopies need to be stored safely to ensure that no other person outside the immediate field team and investigators have access to the data.

If data is collected using paper and pencil, data entry should be done by reliable data clerks and using double entry[16]. Free software programmes can be used to facilitate data entry . For the final preparation of data extensive consistency, and logical checks needs to be performed and ranges of values should be checked. If possible, inconsistent information should be rechecked with interviewers and best be corrected by re-interviewing participants.

Data analysis is much facilitated by available statistical programmes such as STATA ®, SPSS® or others. EPI Info is a free statistical software which provides sufficient statistical options to describe data using frequency tables, or to tabulate information in response to background characteristics such as age, education, place of living or others. Also common statistical analysis can be done using EPI Info.

Frequency tables and tabulation are of utmost importance to understand data and should always be part of the analysis, even when more advanced statistical methods are employed. Histograms (for continuous data such as age) and bar charts (to display categorical data such as marital status) can greatly support the data presentation.

For the presentation of tables it is important that: 1) the title is concise; 2) rows and columns are clearly labelled; 3) totals and the unit of measurement is given, and 4) if needed the source of data is added. Similarly, graphs should have 1) clear labels including units of measurement for x and y axis, 2) scales should start with zero, 3) and graphs should be kept simple.

A plan for archiving, documentation and dissemination should best be prepared already during the study planning phase. A central, protected place to store the original data is needed. Before data can be circulated to other people for further analysis, the datasets should be anonymised, so that essential information which would enable to indentify a participant or interviewers is removed. If data are made available to others, 'meta-data' need to be prepared. Meta data include at least information on study title, where, when and by whom the data collection was done, overall size and the sampling methods, and the instruments used (questionnaires). Best the information on the individual variable (names, labels, categories, codes) should be included in the meta data.

Report needs to be made available within a reasonable time limit which should be well less than one year. Publications in peer reviewed journals typically need much more time and cannot replace reports which are also an important feedback to those that supported the study and contributed to the conduct.

Key questions in field work organisation, data management, data analysis

Is a training manual available and trainers identified?

Is training and capacity building for the data collection team sufficient?

Is a field manager / supervisor part of the team?

[16] http://wwwn.cdc.gov/epiinfo/

Is a continuous quality assurance mechanism been employed?

Are daily summary sheets prepared and values/information checked?

Is a field instruction manual available?

Are daily data back-up's / photocopies done?

Is data entry done according to standard (double entry / quality checks)?

Does the data analysis include clear tables for all major variables?

Are stratified tables prepared to show distribution in response to main background characteristics?

Are simple histograms or bar charts included in the report?

6. Data in emergency settings

Emergency or disaster settings present a particular challenge for getting reliable data for baseline and follow up. Disaster settings are defined as a serious disruption of the functioning of a community or a society causing widespread human, material, economic or environmental losses which exceed the ability of the affected community or society to cope using its own resources. Disaster or emergency setting refer to situation or event, which overwhelms local capacity, necessitating a request to national or international level for external assistance[35].

Whereas using appropriate data for results measurement or monitoring of progress already presents a challenge in many low and middle income countries, this applies even more so to disaster or emergency settings. Here often the total number of population in need is already difficult to estimate due to deaths and population movement. As needs are overwhelming, time investments in more thorough assessment of the situation is often considered unethical in the first response period. Emergency aid is provided by many different, smaller or larger actors. Even if some coordination between aid agencies has become the standard, joint assessments in the immediate disaster situation is not [36].

Thus reliable baseline data are rare in disasters and emergency situations. New technologies such as tracking mobile phones might present an opportunity of at least assessing population movements [37]. In this example, the geographical positioning of mobile phone SIM (subscriber identity module) cards was used to get an idea where the population moved, first after the earthquake and later, after the cholera outbreak in Haiti in 2010. Even if these data have clear limitations - not all people own mobile phones, particularly not the poorest, small children or old people - the pattern found was very similar to a retrospective population-based survey.

Methods much used for baseline assessment during the first days up to the first weeks include use of reports, maps, surveys as well as data from before the onset of the disaster situation and qualitative approaches, such as reports from witnesses or assessments using checklists. Although these data sources present important sources for decision making, they have often shown to be biased [38].

More rigorous assessment using quantitative approaches are typically done after the first days or weeks after the onset of the emergency. Different approaches are used and assessment manuals and tools are available from different agencies [17]. Sampling is a difficult task in emergency settings as population counts are rarely available and sampling frames can hardly even been established. Local key informants such as religious leaders might be able to give better information than government officials [39]. Médecins sans Frontières piloted an assessment at health care delivery points including people that did seek care after the Pakistan earthquake in 2005 to obtain information on the impact on health and living situation. The estimates obtained were relatively similar to later published official statistics thus giving an indication that interviews at health care delivery points might be an option to get relatively good baseline data [40].

More rigorous assessments use baseline data reconstructed from available pre-disaster surveys combined with records and registers, and also by asking people to recall information from before the disaster occurred. Essential here is the use of a broad method mix including qualitative and qualitative data with extensive triangulation. Triangulation described systematic use and comparison of different data sources and derived through different methods. By this validity and in particular ‘generalisability’ can be greatly increased [41].

[17] See a annotated bibliography on Rapid Needs Assessment of health and nutrition in humanitarian emergencies: *http://ki.se/content/1/c6/03/09/69/Annotated%20bibliography%20on%20rapid%20needs%20assessment.pdf*

References

1. Bryce, J., et al., *The Accelerated Child Survival and Development programme in west Africa: a retrospective evaluation.* The Lancet, 2010. **375**(9714): p. 572-582.
2. Victora, C.G., et al., *Measuring impact in the Millennium Development Goal era and beyond: a new approach to large-scale effectiveness evaluations.* The Lancet, 2011. **377**(9759): p. 85-95.
3. Habicht, J.P., C.G. Victora, and J.P. Vaughan, *Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact.* Int. J. Epidemiol., 1999. **28**(1): p. 10-18.
4. Monitoring und Evaluation group of IHP+. *Monitoring Performance and Evaluating Progress in the Scale-up for Better Health. A Proposed Common Framework*. [cited 2011 Sep]; Available from: http://www.internationalhealthpartnership.net//CMS_files/documents/a_proposed_common_framework_EN.pdf.
5. Boerma, J.T., et al. *Monitoring and evaluation of health systems strengthening* 2010; Available from: http://www.internationalhealthpartnership.net//CMS_files/documents/monitoring_and_evaluation_of_health_systems_strengthening_-_an_operational_framework__EN.pdf.
6. Siddiqi, S., et al., *Framework for assessing governance of the health system in developing countries: Gateway to good governance.* Health Policy, 2009. **90**(1): p. 13-25.
7. Hotchkiss, D., et al., *Evaluation of the Performance of Routine Information System Management (PRISM) framework: evidence from Uganda.* BMC Health Services Research, 2010. **10**(1): p. 188.
8. WHO, *Assessing the national health information systems. An assessment tool. Version 4.00*, Health Metrics Network, Editor. 2008: Geneva.
9. WHO, *Everybody business: strengthening health systems to improve health outcomes: WHO's framework for action*. 2007: Geneva.
10. WHO, *Reproductive health indicators. Guidelines for their generation, interpretation and analysis for global monitoring*, WHO. Department of Reproductive Health and Research, Editor. 2006: Geneva.
11. MEASURE. *Guidance for selection and using core indicators for cross-country comparisons of health facility readiness to provide services*. 2007; Available from: http://www.cpc.unc.edu/measure/publications/wp-07-97.
12. Stanton, C.K., et al., *Reliability of data on caesarean sections in developing countries.* Bulletin of the World Health Organization, 2005. **83**: p. 449-455.
13. Machiyama, K. *A Re-examination of recent fertility declines in sub-Saharan Africa.* DHS Working Papers 2010; Available from: http://www.measuredhs.com/pubs/pdf/WP68/WP68.pdf.
14. Lu, C., et al., *Limitations of methods for measuring out-of-pocket and catastrophic private health expenditures.* Bulletin of the World Health Organization, 2009. **87**: p. 238-244D.
15. MEASURE. *Data Quality Audit Tool. Guidelines for Implementation.* 2008; Available from: http://www.cpc.unc.edu/measure/tools/monitoring-evaluation-systems/data-quality-assurance-tools/dqa-auditing-tool-implentation-guidelines.pdf.
16. WHO, *Manual on use of Routine Data Quality Audit (RDQA) tool for TB monitoring*, WHO Stop TB Department, Editor. 2010: Geneva.
17. Hunger, C., et al., *Assessing unmet obstetric need in Mtwara Region, Tanzania.* Tropical Medicine & International Health, 2007. **12**(10): p. 1239-1247.
18. JAMUHURI YA MUUNGANO WA TANZANIA. OFISI YA WAZIRI MKUU. TAWALA ZA MIKOA NA SERIKALI ZA MITAA, *TAARIFA YA HUDUMA ZA AFYA YA UZAZI NA MTOTO KANDA YA KUSINI .MWAKA 2010*. 2010.
19. WHO, *International Classification of Diseases and Related Health Problems. 10th Revision*. 1994, Geneva: WHO
20. Murray, C.J.L., et al., *What Can We Conclude from Death Registration? Improved Methods for Evaluating Completeness.* PLoS Med, 2010. **7**(4): p. e1000262.
21. Mathers, C.D., et al., *Counting the dead and what they died from: an assessment of the global status of cause of death data.* Bulletin of the World Health Organization, 2005. **83**: p. 171-177c.
22. Schuitemaker, N., et al., *Underreporting of Maternal Mortality in The Netherlands.* Obstetrics & Gynecology, 1997. **90**(1): p. 78-82.
23. Schutte, J.M., et al., *Rise in maternal mortality in the Netherlands.* BJOG: An International Journal of Obstetrics & Gynaecology, 2010. **117**(4): p. 399-406.

24. Gissler, M., et al., *Pregnancy-related deaths in four regions of Europe and the United States in 1999-2000: Characterisation of unreported deaths.* European Journal of Obstetrics & Gynecology and Reproductive Biology, 2007. **133**(2): p. 179-185.

25. Daniela, K.-T., et al., *Under-reporting of direct and indirect obstetrical deaths in Austria, 1980-98*. Acta Obstetricia et Gynecologica Scandinavica, 2002. **81**(4): p. 323-327.

26. Hogan, M.C., et al., *Maternal mortality for 181 countries, 1980-2008: a systematic analysis of progress towards Millennium Development Goal 5.* The Lancet, 2010. **375**: p. 1609-1623.

27. WHO, et al., *Trends in Maternal Mortality: 1990 to 2008. Estimates developed by WHO, UNICEF, UNFPA, and The World Bank*. 2010, WHO: Geneva.

28. Cousens, S., et al., *National, regional, and worldwide estimates of stillbirth rates in 2009 with trends since 1995: a systematic analysis.* The Lancet, 2011. **377**(9774): p. 1319-1330.

29. Black, R.E., et al., *Global, regional, and national causes of child mortality in 2008: a systematic analysis.* The Lancet, 2010. **375**(9730): p. 1969-1987.

30. Rajaratnam, J.K., et al., *Neonatal, postneonatal, childhood, and under-5 mortality for 187 countries, 1970-2010: a systematic analysis of progress towards Millennium Development Goal 4.* The Lancet, 2010. **375**(9730): p. 1988-2008.

31. AbouZahr, C., *New estimates of maternal mortality and how to interpret them: choice or confusion?* Reproductive Health Matters, 2011. **19**(37): p. 117-128.

32. Shirima, K., et al., *The use of personal digital assistants for data entry at the point of collection in a large household survey in southern Tanzania.* Emerging Themes in Epidemiology, 2007. **4**(1): p. 5.

33. Paquette, D., J. Bryant, and J. de Wit, *Respondent-Driven Sampling and the Recruitment of People with Small Injecting Networks.* AIDS and Behavior: p. 1-10.

34. Malekinejad, M., et al., *Using Respondent-Driven Sampling Methodology for HIV Biological and Behavioral Surveillance in International Settings: A Systematic Review.* AIDS and Behavior, 2008. **12**(0): p. 105-130.

35. WHO. *Definitions: emergencies.* [cited 2011 5 Sept]; Available from: <u>*http://www.who.int/hac/about/definitions/en/*</u>

36. Buttenheim, A., *Impact evaluation in the post-disaster setting: A conceptual discussion in the context of the 2005 Pakistan earthquake*, in *International Initiative for Impact Evaluation. Working paper 3.* , The International Initiative for Impact Evaluation (3ie), Editor. 2009.

37. Bengtsson, L., et al., *Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti.* PLoS Med, 2011. **8**(8): p. e1001083.

38. van Schreeb, J., *Needs assessment for international humanitarian helth assistance in disasters*. 2007, Karolinska Institutet: Stockholm.

39. Schlecht, J. and S. Casey. *Challenges of collecting baseline data in emergency settings*. RAISE Repoductive Health Access, Information and Service in Emergencies 2008; Available from: *http://www.raiseinitiative.org/projects/introduction.php.*

40. Van Schreeb, J., N. Karlsson, and H. Rosling (2007) *Clinical entrace interviews: a new method to assess needs after a sudden impact disaster*. Open Medicine

41. Bamberger, M., *Strenthening the evaluation of programme effectiveness through reconstructing baseline data.* Journal of Development Effectiveness, 2009. **1**(1): p. 37-59.

42. Pearson, M. *Impact Evaluation of the Sector Wide Approach (SWAp), Malawi. DFID. Human Development Resource Centre & UKAID*. June 2010 [cited 2011 Sep]; Available from: *http://www.dfid.gov.uk/Documents/publications1/hdrc/imp-eval-sect-wde-appr-mw.pdf.*

43. Björkman, M. and J. Svensson, *Power to the people: Evidence from a randomized field experiment on community-based monitoring in Uganda.* The Quarterly Journal of Economics, May, 2009.

44. Arifeen, S.E., et al., *Effect of the Integrated Management of Childhood Illness strategy on childhood mortality and nutrition in a rural area in Bangladesh: a cluster randomised trial.* The Lancet, 2009. **374**(9687): p. 393-403.

Annex

Annotated Bibliography

Frameworks for results measurement and impact evaluation

WHO. Monitoring and evaluation of health system strengthening. Prepared by Boerma T, Abou-Zahr C, Bos E.Hanos, P, Addai, E and Low Beer D. Geneva, 2010

The paper presents an overall framework for monitoring and evaluation of health system strengthening and discusses how it can be operationalized at country level and how global partners can work together. The framework as given in the introduction of this manual is taken from this publication. The publication describes in more detail general principles of usage of this framework, the core indicators, data sources, data analysis and dissemination. This paper presents thus an important overall guidance of how monitoring and evaluation can be done in line with the Paris declaration.

The homepage of the **International Initiative for Impact evaluation** offers a good overview on the debate of best ways to evaluate programmes. Several working papers on how to do impact evaluation are found. *http://www.3ieimpact.org/admin/pdfs_papers/11.pdf*

Study design, surveys, epidemiological methods

Wassenich P. Data for Impact Evaluation. The World Bank. Doing Impact Evaluation No. 6. October 2007. http://siteresources.worldbank.org/INTISPMA/Resources/383704-1146752240884/Doing_ie_series_06.pdf

The document gives a good in overview about important aspects in study design, epidemiological concepts relevant to quantitative studies. The author also includes some information where secondary data can be found, although not specifically in respect to health. Important concepts in household survey design and conduct are covered such as questionnaire design, sampling, quality assurance during field work and data auditing.

International Household Survey Network. Survey Quality Assessment Framework SQAF. Prepared by the Statistical Service Centre of the University of Reading.

http://www.surveynetwork.org/home/index.php?

The publication is an excellent checklist which is of good use while preparing and conducting a household survey. It alerts what to consider in the planning phase such as to check whether all stakeholder are include, staff is complemented, and budgets are realistic and complete. In chapter 4 in this publication a brief and useful overview on the terminology used in sample design and sampling approaches is given. The next chapter gives guidance what to look at in relation to questionnaire design. The Chapters 6 include important aspect of quality assurance in field work organisation and implementation. The chapters 7 to 10 include aspects in data management, data analysis, dissemination and documentation including meta data.

A slide series presented during a congress on impact evaluation: **Mind the Gap** Forum in June 2011 in **Cuernavaca** http://www.impactevaluation2011.org/forum/ gives many good figures and messages to understand key epidemiological principles in quality of data.

http://www.impactevaluation2011.org/forum/wp-content/uploads/2011/06/C-8-file-1.pdf

Dicklberger et al, Baselineerhebung. Ein Leitfaden zur Planung, Durchführung, Auswertung und Nutzung der Ergebnisse, GTZ 2010

The publication gives in particular good advice for the planning phase of a baseline study. It puts the need for monitoring and evaluation in the overall context of the goals of the German Development Cooperation. In the annex good examples are found how to collect data needed (observation, interviews, qualitative design.

Health Care Evaluation (Understanding Public Health) [Paperback]. Sarah Smith, Don Sinclair, Rosalind Raine, Barnaby Reeves. 2005. Open University Press. www.openup.co.uk. 2005. ISBN -10: 0335218490.

A textbook prepared for MSc students. The content covers key epidemiological concepts and goes well through the essential knowledge needed in evaluating health care, the design, evaluating costs and cost-effectiveness as well as equity.

Introduction to Epidemiology (Understanding Public Health) [Paperback]. Bailey L, Vardulaki K, Lanhhan L, Chandramohan D. 2005. Open University Press. www.openup.co.uk.

ISBN -10: 0335214334.

A textbook prepared for MSc students (as above). The context comprises the classical themes in epidemiology, different measurements, association and impact, study types (case control, cohort etc), risk assessments, and surveillance.

Emily White, Bruce K Armstrong, Rodolfo Saracci. Principles of Exposure Measurement in Epidemiology: Collecting, Evaluating and Improving Measures of Disease Risk Factors [Paperback].

2nd edition.Oxford University Press 2008. ISBN978-0-19-850985-1.

This textbook is a very practical guide through essential points in data measurements such as type of data, design of questionnaires and methods to obtain data. The textbook also includes a chapter going in detail into the key quality concepts validity and reliability.

The book is in particular helpful for a critical check of questionnaire design, such as wording and layout.

Assessment of national health information systems

WHO. Health Metrics Network. Assessing the national health information systems. An assessment tool. Version 4.00, 2008

This guidance prepared by WHO leads through an assessment of all different sources of health information in a country, such as census, household surveys, demographic sites, national health information systems (HMIS), registers, financial records. The manual centers less on all aspects of quality in terms of epidemiological soundness but on timely availability, management issues, use of all data sources and bringing them together for an comprehensive assessment of progress. The framework on page 30 is good to have a clear picture of different data sources used and what they represent.

USAID, Data Quality Audit Tool. Guidelines for Implementation, Sep 2008

The guidelines present a complete tool how to audit and assess the data quality of data national service records / HMIS data. The tool guide through the assessment of the overall M&E structures, indicator definitions, data collection and reporting tools, data management process and linkages with national reporting systems. The guidelines include many checklists relevant at different levels (facility, district/aggregation level, national M&E unit). The chapter XX is based on this assessment tool.

Gives also good short introduction into quality standards (validity, reliability,...with examples).

WHO Stop TB department. Manual on use of routine data quality assessment tool for TB monitoring. 2011

The assessment uses steps and aspects as also introduced in the USAID data quality audit tool. Thus it is an assessment of the TB specific routine health service data collection. Several checklists and precise guidance on what to look for are given.

Additional boxes and figures

Box 1: Evaluation designs

Evaluation designs: adequacy, plausibility and probability of programme performance and impact

In assessing public health programmes three types of evaluation called adequacy, **plausibility and probability** have been suggested. They are based on three different types of study design [3, 4].

Adequacy evaluation assesses whether health or behaviour indicators are improving among programme recipients/ populations. Adequacy assessments require no control group. Indicators are compared over time. Cross-sectional or longitudinal studies on single occasion or baseline and repeated measurement to detect trends can be used. Adequacy evaluations are limited to describing whether or not changes have taken place. Improved case management for example, can be safely attributed to a programme on training health workers, but it may be difficult to say that changes in coverage or health outcomes are due to the programme.

Plausibility assessment aims to answer the questions: Did the programme seem to have an effect above and beyond other external influences? Plausibility assessment demands a control group which is mostly chosen in an opportunistic way, for example a neighbouring district (external comparison) or a group of people who were not reached with the intervention (internal comparison). Plausibility assessment can be used to make a careful statement that an intervention or programmes appears to be beneficial or appears to have an influence on change.

Probability assessment aims to answer the question whether the programme/intervention did have a significant effect. Probability assessment demands randomised control group(s) comparing before-after changes between intervention and control group. Although even randomisation cannot guarantee that confounding is eliminated, a randomised design is the basis on which a 'probability' statement can be given, e.g. that intervention A with a defined significance level decreases mortality by 30%. Probability statement can be based on both, individually and community randomized trials.

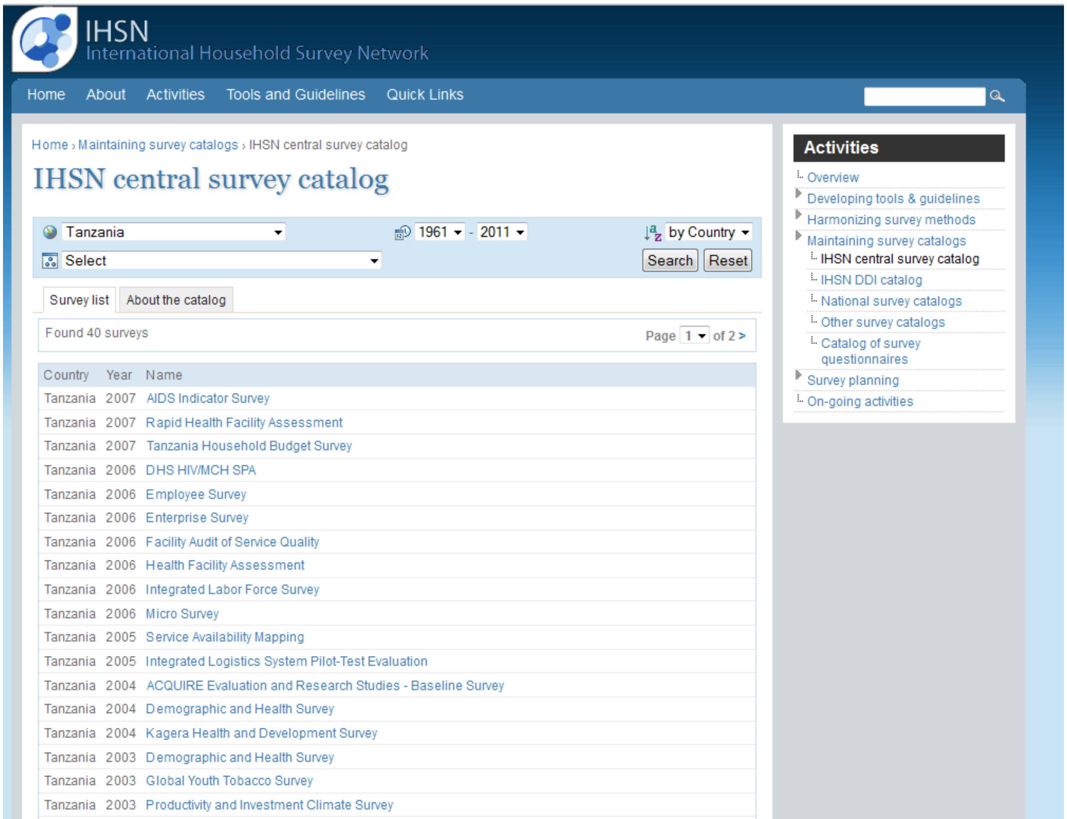
Box 2: Key quality concepts

Dimension of quality	Operational definition and examples
Validity	**Also known as accuracy. The degree to which the data actually measures the intended outcome.** For example, if the outcome to measure is a change (improvement) in health service delivery immunization coverage would be a more valid indicator than household (hh) expenditure on health (since an increase in the latter can also indicate other changes, e.g. due to better socio-economic conditions hh have more resources to spend on health or a deterioration of the health status.
Reliability	**Also called repeatability. The degree to which a test/or question will produce the same results if repeated.** For questionnaire survey reliability is much about how clear a question is formulated (see more chapter 4.2). For some questions used in public health 'intra and inter-observer reliability' has been evaluated extensively and low degrees of reliability are often found in particular for questions which are sensitive such as alcohol intake, or sexual activity. For routine data clear case definitions are of utmost important to improve reliability.
Precision	**The degree how point estimates reflect the true value. Small confidence or uncertainty intervals reflect high precision.** Precision refers to the confidence interval for results obtained in sample surveys which depend on sample size and sampling method (see 4.2). For maternal mortality data uncertainty intervals are given, which represent the sum of all inaccuracy of measurement, including confidence intervals, measurement issues introduced through definition and methods how data are obtained. For data from routine data sources precision depends on the case definition and whether all information needed are collected. If data should be disaggregated by sex, then sex has to be included in the data collection tools.
Completeness	**The degree to which all participants / cases, which were supposed to be included are in fact included in the data collection.** For questionnaire surveys, non-response such as denial to respond to certain questions, poses the problems that a 'bias' might be introduced. Moreover low completeness reduces the 'generalisability' of results (see 4.2b). For routine data collection it is evident that missing information from certain selected persons or units makes summary tables incomplete.
Timeliness	Data should be **timely** available to the people who need the information. Which time lag is acceptable depends primarily on the type of data. For census data a time lag of two years is acceptable as much time is needed for data cleaning and quality checks. For household surveys which also involve extensive quality checks, cleaning of data and data presentation, a time lag up to one year is acceptable. For routine data collection a time lag of one year would be inappropriate, and even more so for routine surveillance of epidemiological outbreaks (such as cholera). Timeliness of health data is in particular a concern of routine health data. Here, the timeliness is often a measure of how effective the overall data management system is.
Integrity	**ntegrity of data** refers to the way data is managed and how results and presentation of results is done. Data management and data protection measures should ensure that nobody can manipulate data after they have been collected, not for scientific, political or personal reasons.

Confidentialtiy	**Confidentiality** refers to anonymity and appropriate data handling. The individual, but also the health facility or unit producing the data should be assured that data and information are not disclosed in an inappropriate way. For household survey data names and other personal information, which would assist to localize an individual should be removed before the data set is handed over to anyone else that the principal investigators.

Figure 8: Snapshot from International Household Survey Network

(Here a list of all surveys ever done in a country can be found)



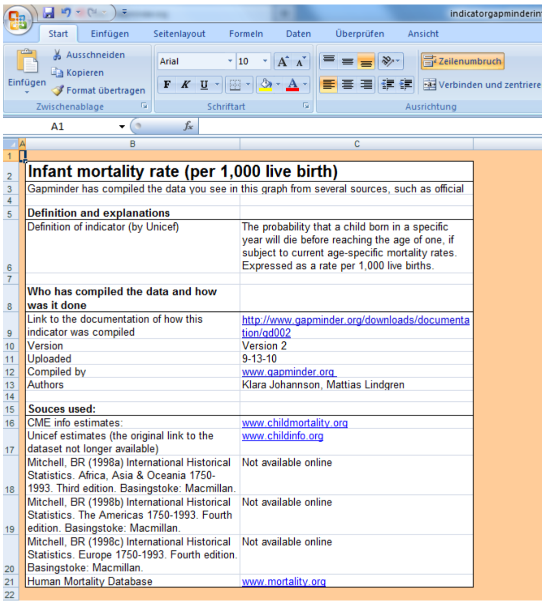
Home › Maintaining survey catalogs › IHSN central survey catalog

IHSN central survey catalog

Found 40 surveys

Country	Year	Name
Tanzania	2007	AIDS Indicator Survey
Tanzania	2007	Rapid Health Facility Assessment
Tanzania	2007	Tanzania Household Budget Survey
Tanzania	2006	DHS HIV/MCH SPA
Tanzania	2006	Employee Survey
Tanzania	2006	Enterprise Survey
Tanzania	2006	Facility Audit of Service Quality
Tanzania	2006	Health Facility Assessment
Tanzania	2006	Integrated Labor Force Survey
Tanzania	2006	Micro Survey
Tanzania	2005	Service Availability Mapping
Tanzania	2005	Integrated Logistics System Pilot-Test Evaluation
Tanzania	2004	ACQUIRE Evaluation and Research Studies - Baseline Survey
Tanzania	2004	Demographic and Health Survey
Tanzania	2004	Kagera Health and Development Survey
Tanzania	2003	Demographic and Health Survey
Tanzania	2003	Global Youth Tobacco Survey
Tanzania	2003	Productivity and Investment Climate Survey

Figure 9: Example of meta data available (from www.Gapminder.org)



	B	C
1		
2	**Infant mortality rate (per 1,000 live birth)**	
3	Gapminder has compiled the data you see in this graph from several sources, such as official	
4		
5	**Definition and explanations**	
6	Definition of indicator (by Unicef)	The probability that a child born in a specific year will die before reaching the age of one, if subject to current age-specific mortality rates. Expressed as a rate per 1,000 live births.
7		
8	**Who has compiled the data and how was it done**	
9	Link to the documentation of how this indicator was compiled	http://www.gapminder.org/downloads/documentation/gd002
10	Version	Version 2
11	Uploaded	9-13-10
12	Compiled by	www.gapminder.org
13	Authors	Klara Johannson, Mattias Lindgren
14		
15	**Souces used:**	
16	CME info estimates:	www.childmortality.org
17	Unicef estimates (the original link to the dataset not longer available)	www.childinfo.org
18	Mitchell, BR (1998a) International Historical Statistics. Africa, Asia & Oceania 1750-1993. Third edition. Basingstoke: Macmillan.	Not available online
19	Mitchell, BR (1998b) International Historical Statistics. The Americas 1750-1993. Fourth edition. Basingstoke: Macmillan.	Not available online
20	Mitchell, BR (1998c) International Historical Statistics. Europe 1750-1993. Fourth edition. Basingstoke: Macmillan.	Not available online
21	Human Mortality Database	www.mortality.org
22		

Published by

Deutsche Gesellschaft für
Internationale Zusammenarbeit (GIZ) GmbH

Programme to Foster Innovation, Learning and Evidence in HIV and Health Programmes of German Development Cooperation (PROFILE),
Component "Results measurement and applied health research"

Registered offices
Bonn and Eschborn, Germany

Friedrich-Ebert-Allee 40
53113 Bonn, Germany
T +49 228 44 60-0
F +49 228 44 60-17 66

Dag-Hammarskjöld-Weg1-5
65760 Eschborn, Germany
T +49 6196 79 -1927
F +49 61 96 79- 801927

Stefan.Weinmann@giz.de
www.giz.de

Authors
Claudia Hanson, evaplan GmbH International Health at the University Hospital Heidelberg
Siegrid Tautz, evaplan GmbH International Health at the University Hospital Heidelberg

Contributions
Stefan Weinmann, GIZ
Jasmin Dirinpur, GIZ

As at: 01/2012
GIZ is responsible for the content of this publication.

On behalf of
Federal Ministry for Economic Cooperation
and Development (BMZ)

BMZ Bonn
Dahlmannstraße 4
53113 Bonn
Germany
T +49 228 99 535-0
F +49 228 99 535-3500

BMZ Berlin | im Europahaus
Stresemannstraße 94
10963 Berlin
Germany
T +49 30 18 535-0
F +49 30 18 535-2501

poststelle@bmz.bund.de
www.bmz.de