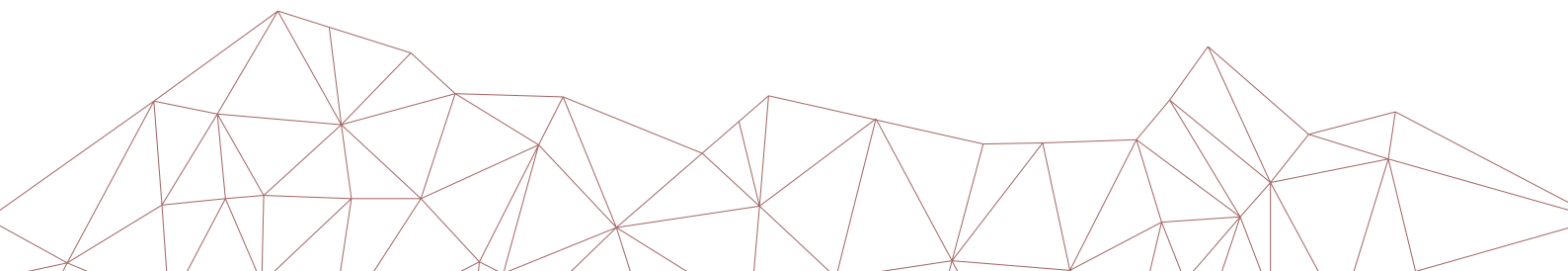




# CRVS technical guide

Guidance for assessing and interpreting the quality of mortality data using ANACONDA

October 2017





## Resources available from the University of Melbourne, Bloomberg Philanthropies Data for Health Initiative

### *CRVS course prospectuses*

These resources outline the context, training approach, course content and course objectives for the suite of CRVS trainings delivered through the Bloomberg Philanthropies Data for Health Initiative. Each course focuses on a specific CRVS intervention or concept, and is designed to support countries to strengthen their CRVS systems and data.

### *CRVS Fellowship reports and profiles*

The CRVS Fellowship Program aims to build technical capacity in both individuals and institutions to enhance the quality, sustainability and health policy utility of CRVS systems in Fellows' home countries. *Fellowship reports* are written by Fellows as a component of the program, and document, in detail, the research outcomes of their Fellowship. *Fellowship profiles* provide a summary of Fellows' country context in relation to CRVS, an overview of the Fellowship experiences, the research topic and the projected impact of findings.

### *CRVS analyses and evaluations*

These analytical and evaluative resources, generated through the Initiative, form a concise and accessible knowledge-base of outcomes and lessons learnt from CRVS initiatives and interventions. They report on works in progress, particularly for large or complex technical initiatives, and on specific components of projects that may be of more immediate relevance to stakeholders. These resources have a strong empirical focus, and are intended to provide evidence to assist planning and monitoring of in-country CRVS technical initiatives and other projects

### *CRVS best-practice and advocacy*

Generated through the Initiative, CRVS best-practice and advocacy resources are based on a combination of technical knowledge, country experiences and scientific literature. These resources are intended to stimulate debate and ideas for in-country CRVS policy, planning, and capacity building, and promote the adoption of best-practice to strengthen CRVS systems worldwide.

### *CRVS country reports*

CRVS country reports describe the capacity-building experiences and successes of strengthening CRVS systems in partner countries. These resources describe the state of CRVS systems-improvement and lessons learnt, and provide a baseline for comparison over time and between countries.

### *CRVS technical guides*

Specific, technical and instructive resources in the form of *quick reference guides*, *user guides* and *action guides*. These guides provide a succinct overview and/or instructions for the implementation or operation of a specific CRVS-related intervention or tool.

### *CRVS tools*

Interactive and practical resources designed to influence and align CRVS processes with established international or best-practice standards. These resources, which are used extensively in the Initiative's training courses, aim to change practice and ensure countries benefit from such changes by developing critical CRVS capacity among technical officers and ministries.

Published by the University of Melbourne, Civil Registration and Vital Statistics Improvement, Bloomberg Philanthropies Data for Health Initiative.

Melbourne School of Population and Global Health  
Building 379  
207 Bouverie Street  
Carlton, VIC 3053  
Australia

CRVS-info@unimelb.edu.au

[www.mspgh.unimelb.edu.au/dataforhealth](http://www.mspgh.unimelb.edu.au/dataforhealth)

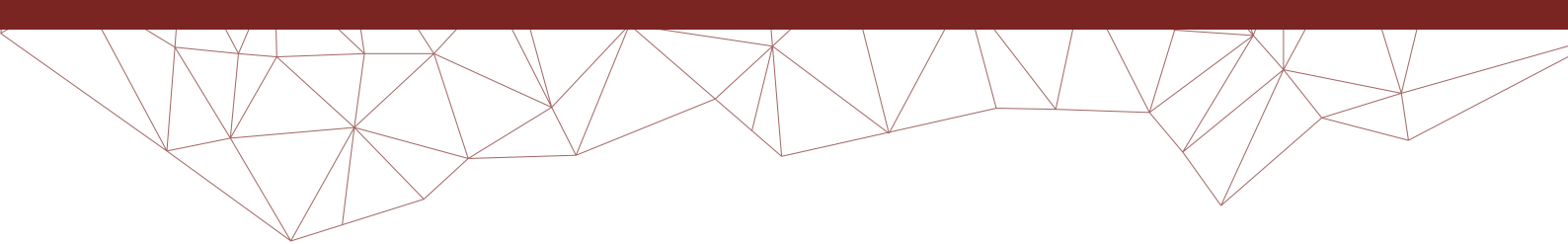
**Made possible through funding from Bloomberg Philanthropies**  
[www.bloomberg.org](http://www.bloomberg.org)

### **Suggested citation**

Mikkelsen, L, Lopez, A. *Guidance for assessing and interpreting the quality of mortality data using ANACONDA*. CRVS technical guides. Melbourne, Australia: Bloomberg Philanthropies Data for Health Initiative, Civil Registration and Vital Statistics Improvement, University of Melbourne; 2017.

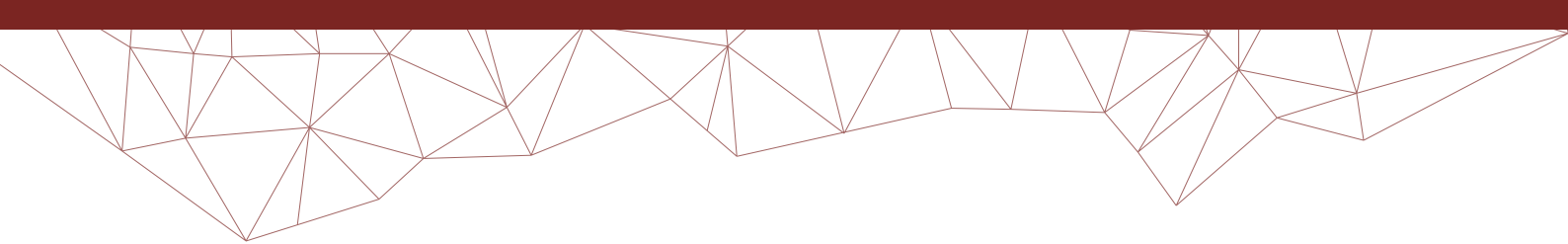
### **Acknowledgements**

The authors are grateful to Carla AbouZahr, Tim Adair and Nicola Richards, University of Melbourne; Michael Hegnauer, University of Basel; and Martin Bratschi and Philip Setel, Vital Strategies, who reviewed the document; and to Doris Mafat, World Health Organization, and Don de Savigny, University of Basel, for their helpful comments on an earlier version.



## Contents

Acronyms and abbreviations	iv
About this document	1
Background	2
What is ANACONDA?	2
Why use ANACONDA?	3
The 10 steps of ANACONDA	4
Overview	4
(1) Data inputs and general background checks	5
Step 1: Basic tabulations of deaths by age, sex and cause of death	5
(2) Mortality data	9
Step 2: Crude death rates	9
Step 3: Age-specific mortality rates	12
Step 4: Age distribution of deaths	16
Step 5: Child mortality rates	17
(3) Cause of death analysis	19
Step 6: Classification of deaths into broad cause groups	19
Step 7: Quality of cause of death data	22
Step 8: Age pattern of mortality for broad disease groups	28
Step 9: Leading causes of death	30
(4) Overall data quality index: VSPI(Q)	32
Step 10: Vital statistics performance index	32
Summary	43
ANNEX 1: Infographic of ANACONDA's 10 steps	34
ANNEX 2: Model for estimating completeness	35
References	38



## Acronyms and abbreviations

${}_5q_0$	risk of dying before age 5, per 1000 live births
ANACoD	Analysing mortality levels and Causes of Death
ANACONDA	Analysis of Causes of National Deaths for Action
ASMR	age-specific mortality rate
CDR	crude death rate
COD	cause of death
CRVS	civil registration and vital statistics
GBD	Global Burden of Disease
HIV/AIDS	human immunodeficiency virus/acquired immunodeficiency syndrome
ICD	International Classification of Diseases
ICD 10	International Classification of Diseases 10th edition
IGME	Inter-agency Group for Child Mortality Estimation
IHME	Institute of Health Metrics and Evaluation
UCOD	underlying cause of death
VSPI(Q)	Vital Statistics Performance Index for Data Quality
WHO	World Health Organization



## About this document

The document is written for users or potential users of ANACONDA (Analysis of Causes of National Deaths for Action), an electronic tool for checking the quality of mortality data. Such users are likely to be statisticians and/or analysts in health and statistics departments, researchers, or other experts working with mortality data. Since the first version of ANACONDA was piloted in 2016, many improvements have been implemented. This guide refers to version 3.0 of the application. Although it is likely that further minor development of the tool will happen, the information provided here will remain relevant.

This guidance aims to help users interpret the results of ANACONDA correctly, which may not be easy if they are unfamiliar with mortality analysis. The purpose of each of the 10 steps is made clear, and explanations are offered on how to interpret the outputs, and what in the data to check and verify. The corrective actions that can be used to correct the data and improve their value to policy development are outlined.

A second objective of the guide is to educate users in the principles and approaches of basic mortality analysis. They can then better understand how to critically appraise such data and identify common errors in the data. This also builds knowledge about the patterns one expects to observe in mortality data according to age, sex and cause of death (COD). Careful application of the tool will help to build confidence and skills in the interpretation of mortality statistics, and help people avoid making incorrect judgements about the policy value of the data. Why it is important to carry out systematic quality assessments of mortality data should be obvious after having read this guide.

This document does not provide instructions about how to use ANACONDA; for that, we recommend the *ANACONDA user guide* (University of Melbourne & Swiss Tropical and Public Health Institute 2017), which provides detailed guidance on getting started and running the application. Although not written specifically for those who intend to become trainers in ANACONDA, this guide provides important insight into the logic and intent of the various steps in ANACONDA. Additional background information and guidance to support those who will teach ANACONDA can be found in *ANACONDA: Facilitator's guide* (University of Melbourne 2017).



## Background

The ANACONDA tool was jointly developed by the Melbourne School of Population and Global Health at the University of Melbourne and the Swiss Tropical and Public Health Institute at the University of Basel. It is part of a series of innovative tools developed to improve vital statistics data from civil registration, funded by the Bloomberg Philanthropies Data for Health Initiative at the University of Melbourne.

The basis of ANACONDA is the 10 data quality assessment principles first published by the Health Information Systems Knowledge Hub at the University of Queensland (AbouZahr et al 2010). These principles were subsequently made into an Excel tool named ANACoD (Analysing mortality levels and Causes of Death) by the World Health Organization (WHO).<sup>1</sup>

ANACONDA differs from ANACoD in a number of important ways with regard to both content and technology. In particular, it:

- is built on more flexible and reliable software technology (Java/JavaFX) that makes use of open standards and libraries
- is based on robust software architecture and design principles such as separation of model and view, allowing for easy extension of the algorithms and/or the user interface
- uses an architecture that allows easy deployment to various platforms by providing native packaging and installers, including all necessary software prerequisites
- vastly expands on the set of data quality assessment principles included in ANACoD, especially for the assessment of ill-defined and otherwise incorrect or poorly specified International Classification of Diseases (ICD) codes
- provides a detailed analysis of the CODs that should not be used and are useless for public health analysis
- allows for more flexible data input formats
- has more graphs and charts, and includes an export function for these
- uses more detailed and comprehensive country-comparator information to assess plausibility
- includes more standard demographic and health indicators
- includes an overall summary index of mortality data quality – the Vital Statistics Performance Index for Data Quality, or VSPI(Q), adapted from Phillips et al (2014)
- has an inbuilt data review function where users can enter free text comments and assign an overall status for each step. This function also allows users to produce a data quality report at the end of the process.

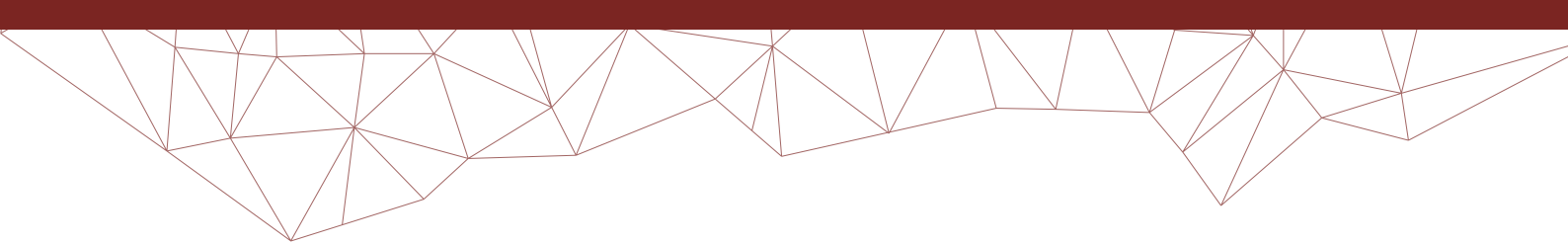
### What is ANACONDA?

The optimal source of mortality and COD statistics for a population is a functioning civil registration system that registers all deaths in the country, with all deaths having a medically certified underlying cause of death (UCOD). ANACONDA is an easy-to-use electronic tool specifically designed to help users analyse the quality of such routine mortality datasets, to understand whether the data are fit for the intended purpose.

ANACONDA analyses the quality of mortality and COD data. It requires only basic computer skills, including familiarity with Microsoft Excel. ANACONDA performs the calculations needed for a comprehensive data quality review, and automatically generates the associated figures and tables from which a data quality assessment report can be written. However, because ANACONDA does not interpret the analyses, it is helpful to have read this guide, particularly if users are not familiar with mortality analysis, ICD and Global Burden of Disease (GBD) classifications.

The structure of the tool is logical, and all the computational steps are automated and straightforward. ANACONDA starts with a broad overview of the input data, applies some simple checks to the mortality data, followed by a detailed assessment of the quality of COD data, and finally computes the VSPI(Q) – the overall index of mortality data quality.

<sup>1</sup> [www.who.int/healthinfo/anacod/en/](http://www.who.int/healthinfo/anacod/en/)



ANACONDA is specifically designed to be applied to large datasets such as those from civil registries. Even so, it can also be used to analyse hospital data, data from longitudinal demographic surveillance sites, or any other data source that routinely collects and generates ICD COD data classified by age and sex, including data from verbal autopsies. To benefit from all the checks and assessments performed in the first five steps, data for the total population at risk, classified by age and sex, need to be provided. If only hospital data are available, the exact population at risk, also referred to as the denominator, can usually not be provided.

Although the application of ANACONDA to individual hospital datasets is possible, it may be necessary to combine several years of data to assemble a dataset with sufficient numbers of deaths to provide stable distributions for the top 20 or so CODs. At least 5 deaths per cause per age group (100 for each age group) are required to avoid distortions in the output from ANACONDA. Optimal datasets are those greater than 2000 deaths each for males and females (4000 total). Application of ANACONDA to datasets with significantly fewer deaths might lead to misleading outputs.

Note that ANACONDA will not by itself improve the quality of the input data; it is designed to *identify problems* that need to be addressed to improve the value of mortality data for guiding health policies and practices. However, it is expected that users who read this guide and apply the tool will build the competence and confidence to critically appraise routine mortality data, from whatever source, and that they will be able to identify the intervention strategies necessary to improve them.

## Why use ANACONDA?

All countries need accurate and up-to-date mortality data for a variety of purposes, including:

- informing health and social policy debates
- monitoring progress relative to national and global development goals
- monitoring trends in diseases and injuries
- evaluating policies designed to improve health outcomes.

However, in many countries, the systems that produce mortality and COD data either do not exist or are poorly developed. As a result, the data they produce are often not reliable enough to be used for the purposes listed above.

As a first step to improving these systems, it is important to gain detailed understanding of the problems with the data, particularly with data completeness and diagnostic accuracy. A common concern with any mortality data produced from civil registration systems is how reliable they are in describing the actual mortality patterns in the population to which they refer. For example, even if a dataset includes all deaths in hospitals, it is important to remember that they are different from deaths that occur in the community and, hence, the data cannot be considered to represent the national mortality situation (Murray et al 2007).

COD data that have been collected – often at great expense – ought to be of sufficient quality and representativeness to be used to their full potential. Poor quality and unrepresentative data can lead to poor health policy decisions and lost opportunities to improve population health.

ANACONDA is particularly useful to those who are responsible for *routine* mortality data collections (eg a civil registration and vital statistics (CRVS) system, a health and demographic surveillance site or a Sample Registration System (SRS) in which verbal autopsy is assessed by physicians to determine the probable COD), because it allows them to monitor the quality of their datasets and pinpoint probable areas of weakness each year. When applied each year monitoring data quality with ANACONDA will allow analysts and policy-makers to make greater use of existing, flawed data by understanding the probable biases. It will also provide critically important intelligence to guide strategies and interventions designed to improve the mortality data system.

# The 10 steps of ANACONDA

## Overview

ANACONDA is designed around **10 steps**, some with numerous substeps, that users are advised to follow in sequence (see Annex 1). These 10 steps can be grouped into four broad categories:

1. data inputs and general background checks (**step 1**)
2. mortality data (**steps 2–5**)
3. COD data (**steps 6–9**)
4. overall data quality index: VSPI(Q) (**step 10**).

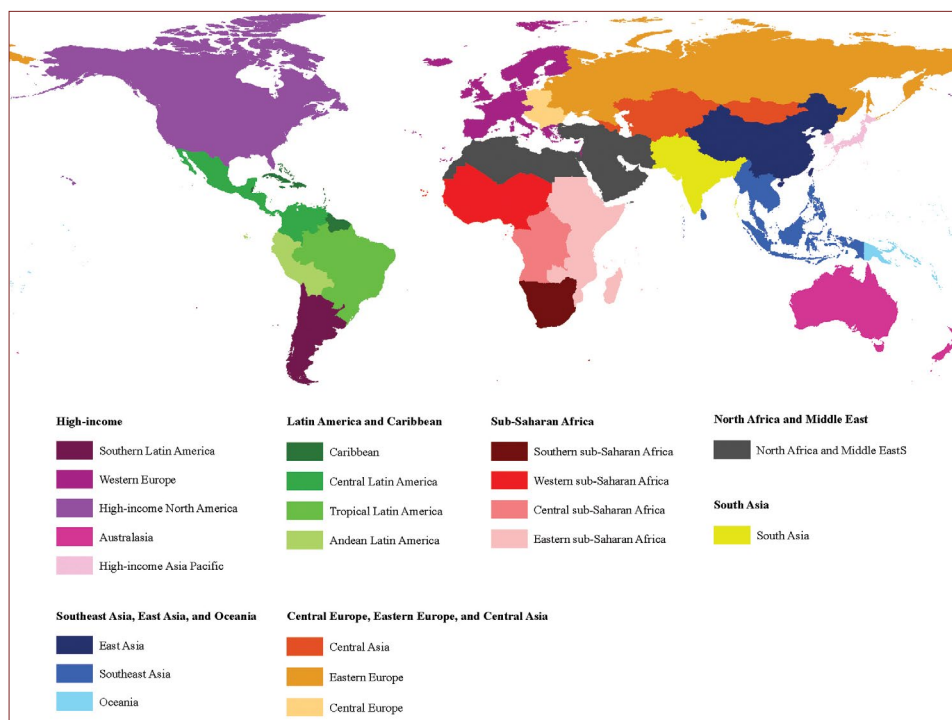
To diagnose possible problems in the mortality input dataset, ANACONDA:

- tabulates and/or graphs the input data in different ways to assess the plausibility of the data based on fundamental demographic and epidemiological relationships
- calculates the proportion and type of unusable COD codes that are of limited or no value for public health analysis
- compares the input data (or some summary measure of them) with a global source of demographic and epidemiological estimates for the country, or geographic region, to assess consistency
- allows users to monitor annual changes in the quality of the dataset through VSPI(Q).

To assess plausibility, the national mortality data for a country are compared with the most recent estimates for that country or a neighbouring region. Most of the comparators are from the GBD study prepared by the Institute for Health Metrics and Evaluation (IHME) in Seattle. Since these are estimates, only major differences should be investigated to check plausibility of input data.

The 21 epidemiological and geographical regions used for comparative purposes in ANACONDA are those defined for the GBD study (Figure 1).

Figure 1: Global Burden of Disease epidemiological regions



Source: Lozano et al (2013)



## (1) Data inputs and general background checks

# 1



### Step 1: Basic tabulations of deaths by age, sex and cause of death

The first assessment step in the tool has eight substeps. These are all intended to check the consistency of the input data and identify any significant errors in the dataset. *It is strongly recommended that these be investigated and corrected before applying the other steps.*

**Substeps 1.1 and 1.2** are introductory and remind users about the sources of their input data. The input data should consist of individual death records aggregated by age (using standard 5-year age groups; see below), sex and COD (using ICD 10 codes at three or more digits).

The input tabulation of mortality data should include, for each ICD code:

- the number of deaths for a specified year, or years, where it is necessary to aggregate data from two or more years to achieve the minimum number of cases (4000) recommended for application of the tool
- the number of deaths by sex – in other words, for males and females separately
- the number of deaths by age at death, using the following standard groupings:
  - those born alive but who died before their first birthday – age in completed months, 0–11
  - those who died aged 1–4 years – age in completed years
  - those who died aged 5–9 years
  - those who died aged 10–14, 15–19, and so on, in 5-year age groups up to 80–84
  - those who died aged 85 and over.

As explained in the user guide (University of Melbourne & Swiss Tropical and Public Health Institute 2017), ANACONDA can only be applied if the input mortality data contain the degree of age specificity defined above. If the terminal age group used is younger than 85+ (eg 65+), it is recommended that users first redistribute deaths in the terminal age group to estimate the likely number of deaths in each 5-year age group up to 85+. This will need to be done for each cause listed in the dataset, and separately for males and females. A model to do this can be derived from the age-cause distributions of deaths estimated for the GBD study available on the IHME website.<sup>2</sup>

It is *extremely important that deaths are separately tabulated by the above indicated age groupings up to 85+* (Box 1). In populations where most deaths occur among the elderly, higher age groups such as 100+ may be encountered. Lower terminal age groups such as 65+ may have been appropriate for mortality tabulations 100 years ago, when comparatively few people lived longer than 65 years, however, that is not the case today, when the average life expectancy is 70 or above for all regions in the world, except the least developed countries in Africa (United Nations 2015). As such, mortality data should be tabulated accordingly. If countries are already tabulating data to ages beyond 85, data binning will automatically be applied at data entry and the data collapsed so that 85+ will be the terminal age group.

The input population data should be the midyear population for the same year as the mortality data, by sex and age group. The mid-period population should be used if the mortality data cover more than 1 year (eg, 30 June 2014 if the calendar years 2013–15 are used). Population estimates are generally available from the decennial census and intercensal projections produced by the national statistics office of the country in question. These data will be used for the calculation of rates and ratios required in **steps 2–5**. Tabulations for males and females will be analysed separately.

<sup>2</sup> <https://vizhub.healthdata.org/mortality/age-estimation>

### Box 1: Reporting age at death

It is important that age at death be recorded with precision. A death occurring to a child aged 4 years and 11 months should be classified in the 1–4 age group. Only when the child has completed the fifth year of life (ie had its fifth birthday) should the death be counted in the 5–9 age group. It is usual practice to use 5-year age groups except for deaths occurring in children under 5, which are subdivided into those occurring within the first year of life (often separately for the first month of life, when risk of death is at its highest) and those occurring in the range 1–4, given the very different risks of mortality at various childhood ages before 5.

Precision is also important at older ages, which should continue to be grouped into 5-year categories at least up to the age of 85.

It is poor practice to only tabulate age of death to some relatively low terminal age such as 55+ or 65+. Increasingly, deaths are occurring after about age 60, and it is extremely important for preventive efforts to distinguish between a death at age 80–84 and an adult death at a much younger age, such as 60–64 or 65–69.

The use of these standard 5-year age groupings is also important because the same age groups are used to compile census data on population size and distribution, which are used by ANACONDA as denominators for the calculation of rates and ratios.

As a side note, because of the importance of accurate data on the age at death, it is recommended to include the date of birth, as opposed to just age, in the data collected about each deceased individual.

**Substeps 1.3 and 1.4** tabulate the input population and mortality data in the standard 5-year age groups and show this graphically as population pyramids. In cases where the source is hospital data, it is not possible to enter any correct corresponding population data and so the steps requiring denominators cannot be performed. Both substeps tabulate the data and display them as population pyramids to allow users to visually inspect their input data for obvious irregularities – for example, large differences between adjacent age groups, gross differences in sex ratios, or deviations from expected male and female patterns.

Understanding the local demographic and mortality situation is important for assessing the plausibility of the tabulations and graphs produced by the steps that follow.

A population pyramid typically consists of two back-to-back bar graphs, with age groups on the vertical axis and population size on the horizontal axis, shown as percentages of the total (Figure 2). The death pyramid for a population is similarly constructed, but with the percentages of deaths at each age shown on the horizontal axis (Figure 3).

In most developed countries, the age–sex pyramid is constructed based on annual birth and death data from the civil registration system and decennial censuses. In countries where civil registration systems are weak, age–sex population pyramids can only be reliably estimated from the census. Intercensal estimates of population size by age and sex generally need to be obtained from mortality rates derived from model life tables, which are inherently uncertain. The United Nations Population Division generates regular updates on national population sex and age structures, which should be used when there is doubt about the reliability of country population data (United Nations 2015).

Figure 2: Example population pyramid

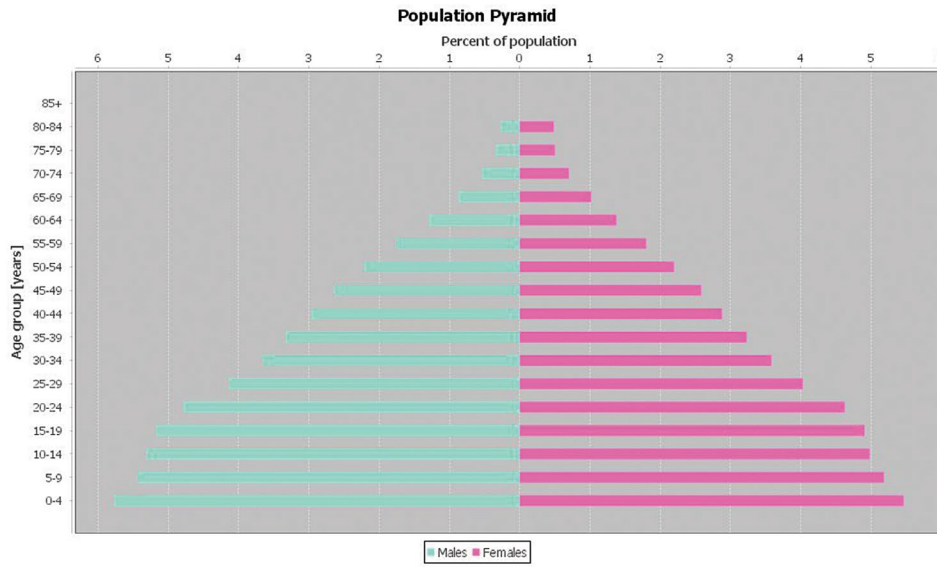
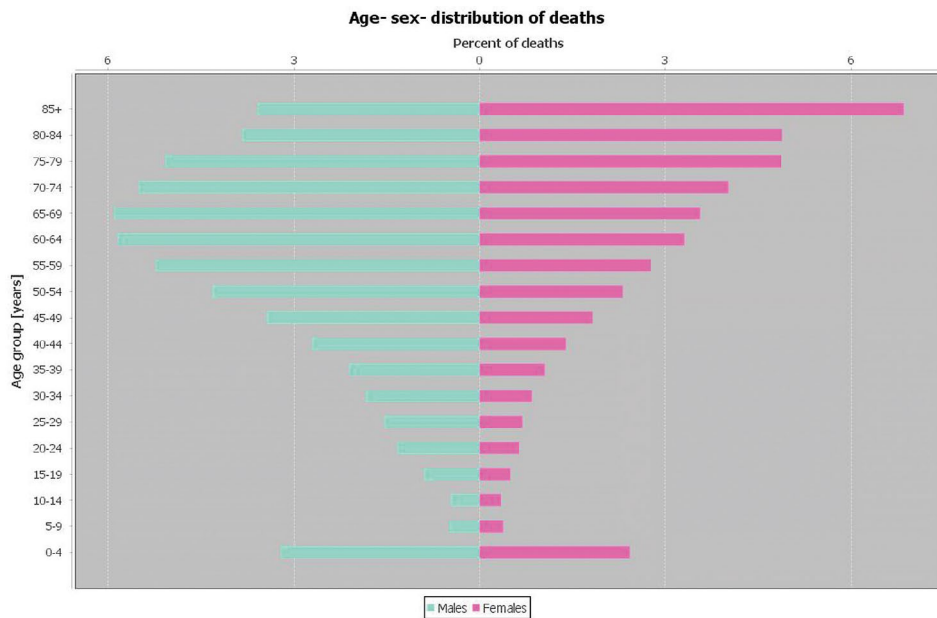


Figure 3: Example death pyramid





### Understanding substeps 1.3 and 1.4

To interpret the two pyramids, it is helpful to do some basic visual checks.

- For the population data, is there any age/sex group that accounts for more than 10% of the population? If so, verify that the input data are indeed correct, or provide an explanation, since such proportions are highly unusual.
- By overlaying them, compare the pyramid calculated from the input data with one based on the estimated population age–sex structure prepared by the United Nations Population Division for the same year or closest match. Major deviations should be investigated. It is important to note that, in all comparisons offered in ANACONDA, deviations from a standard that is being used to evaluate the input data do not necessarily mean that the input data are incorrect. These global standard mortality and population estimates have been prepared using substantial epidemiological and demographic expertise, and large collections of international population and mortality data from which expected patterns can, and have, been derived. However, these are estimates and may not, in certain cases, reflect the reality of a country's input data. Such local departures from a comparative standard are therefore possible, but unlikely, and should be carefully investigated for plausibility.
- The mortality input data are shown as an inverted pyramid, and can also be checked visually. Typically, the pyramid will show fewer deaths at younger ages (except for children aged 0–4) with progressively increasing numbers into old age. It is not unusual for there to be a high proportion of deaths at age 85+ or at the highest age group.
- More males die at younger ages than females, with the reverse being the case at the oldest ages. The death pyramid should reflect this.

**Substep 1.5** shows the input COD data by disease, for age groups 0–4 and 5+. Apart from being useful for various types of analyses, given the different disease patterns for children and adults, this also facilitates the identification of any obvious errors in the data.



### Understanding substep 1.5

Inspection of the COD patterns of children and adults separately, and ranking these, is a quick way of identifying leading diseases in either group or identifying improbable cases. These may include suicides, pregnancies or other adult conditions attributed to children, or neonatal conditions in the adult population. These should all be corrected before proceeding with the subsequent steps.

**Substep 1.6** further refines the age breakdown by compiling the data into broad age groups likely to be of interest for various health promotion, and disease and injury-control programs (eg reproductive health, child health and adolescent health). These age groups have been defined as follows: 0–4, 5–9, 10–19, 20–24, 25–29, 30–49, 50–69 and 70+.<sup>3</sup>

This substep groups the input COD data into a shorter list to facilitate overview and plausibility assessment. For this tabulation, ANACONDA uses the ICD 10 tabulation list of 113 aggregated disease and injury categories (United Nations 2015), intended to include all causes likely to be of public health importance. It is a list that is commonly used for ranking mortality data to guide public health efforts to address the leading CODs. Stratifying the COD data in age groups facilitates understanding and reporting of the leading causes at different stages of life, which are childhood, adolescence, young adulthood, adulthood and old age.



## Understanding substep 1.6

Age remains the most important predictor of the risk of death, with epidemiological patterns of disease at different ages well established from years of observation. As a result, knowing the age structure of mortality from different diseases and conditions is not only important for prevention (eg, adolescent maternal deaths, suicides in young adults), but also a way to check the plausibility of the data given the strong age-dependency observed for most major CODs. For example, the vast majority of lung cancer deaths should occur at older ages. If the input data tabulation shows a different pattern, there are likely to be major problems with misdiagnosis (could these be asthma deaths, which are more common at younger ages?) or underreporting of lung cancer deaths among older adults.

Users may also prefer to tabulate their input data according to the GBD COD list of some 300 causes (Lozano et al 2013). This list is used in **substep 1.7**, which continues to investigate the relationship between age and disease patterns.

**Substep 1.8** is purely arithmetical, and points out if there are any internal inconsistencies in the data.

## (2) Mortality data

# 2



### Step 2: Crude death rates

The second step calculates the crude death rate (CDR) for males and females, based on the input data. As a mortality indicator, the CDR is the simplest measure of population health status (Box 2). ANACONDA also uses the CDR as a simple measure of how complete (in terms of deaths captured by the system) the mortality data are likely to be. In other words, the proportion of deaths are actually reported to authorities. This proportion is known as the ‘completeness’ of registration of deaths in a population and is perhaps the most fundamental component of data quality (Philips et al 2014). Input data with a significant proportion of unreported or unregistered deaths are likely to be substantially biased (towards mortality patterns in the better-off/urban sectors of the population, who are more likely to register deaths), and hence are of reduced value for policy.

### Box 2: Crude death rate

The crude death rate (CDR) is a measure of the number of deaths in a population relative to the size of that population during a given period, usually 1 year. It is expressed in units of deaths per 1000 individuals per year. Thus, a CDR of 9 per 1000 means that, in a population of 1 million, one would expect to see 9000 deaths recorded per year – if all deaths get registered. Both the numerator (9000 in this case) and denominator (1 million) should refer to the same population in terms of geography and time. Also, it is standard practice to use the midyear population size as the best estimate for the average population exposed to the risk of dying over the course of the year.

The CDR is called a ‘crude’ rate because it does not take into consideration the age structure of the population. In practice, the risk of death varies according to age and sex as well as socioeconomic status, environment and other factors. For example, populations with a high proportion of young children or a high proportion of elderly people will, other things being equal, have relatively higher CDRs. This is because mortality risks are highest at youngest and oldest ages. Similarly, mortality rates are generally higher among males than females. Therefore, when comparing populations across countries, geographic areas or over time, it is important to use age- and sex-specific mortality rates alongside the CDR (see step 3).



## Understanding step 2

The CDR always needs to be interpreted in relation to the age–sex structure of the population, shown in substep 1.3. Populations with higher CDRs are likely to be those with high proportions of young children or elderly people, the two stages of life with the highest mortality risks. For example, the CDRs for the two pyramids shown in Figures 4 and 5 are 13 per 1000 (top) and 6 per 1000 (bottom). This difference reflects the fact that countries with ageing populations have small cohorts with the lowest death rates and large cohorts with high death rates (ages 60+). The reverse is likely to be true in countries with younger populations.

Based on many decades of observing CDRs, demographers have concluded that there is generally a lower limit for the CDR of around 5 per 1000. For example, during the past 20–30 years, Japan, a country with high life expectancy, has consistently registered the lowest age-specific mortality rates in the world. Yet throughout this period, the CDR in Japan never fell below 5 per 1000. Thus, any CDR below 5 per 1000 should be viewed with caution, since it is likely to reflect an underregistration of deaths rather than a true reflection of low overall mortality rates. Key checks to perform include assessing the age structure of the population via the population pyramid (to understand whether there are significant proportions of elderly or young children in the population, which will then affect overall death rates), checking that the CDR is higher for males than females and comparing it with the IHME-estimated CDR, and examining the time trend in the CDR.

There are exceptions. In countries where child mortality rates have fallen rapidly but too recently to result in a significant increase in the number of elderly, the population pyramid is heavily biased towards people in the young adult age range, and the high mortality age groups remain small. Examples are mostly found among the Gulf States, and some counties in Latin America, which all have low CDRs of around 3–4 per 1000.

The CDR should be calculated separately for males and females and should be slightly higher for males; if not, it is likely that male deaths are underreported.

To add additional perspective on the plausibility of the calculated CDR, ANACONDA offers a comparator in the form of the CDR calculated by IHME. If the difference between the calculated CDR and that estimated for the comparator is large (3 points or more), it is likely a reflection of underregistration of deaths by the civil registration system. From the IHME GBD database, ANACONDA calculates a CDR trend line going back to 1990 showing the evolution of mortality in the given country in the past 35 years. Mortality levels change slowly with only small fluctuations from year to year; hence, large fluctuations in the CDR calculated from the input data should be investigated to make sure that they can be explained – perhaps by some sudden catastrophe.

Since the CDR is calculated from the number of deaths occurring in the population at risk (usually the midyear resident population), it is not relevant for hospital data or other data for which the population at risk is not known.

Overall, the CDR calculated from the input data is a powerful and simple test of the likely extent of underregistration of deaths by the civil registration system. Any CDR of less than 5 per 1000 population needs to be critically assessed for plausibility, as it most likely is due to underregistration of deaths.

Figure 4: Example of an ageing population pyramid (CDR of 7 per 1000 population)

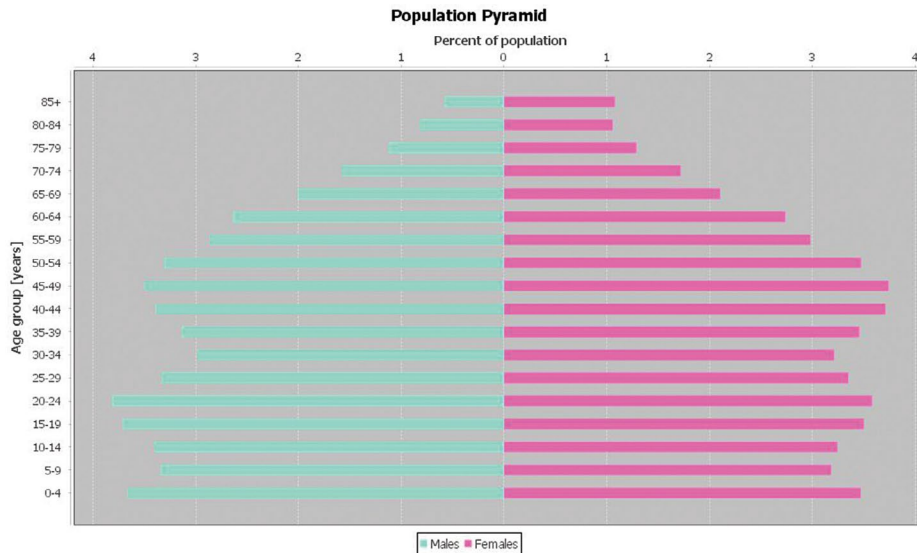
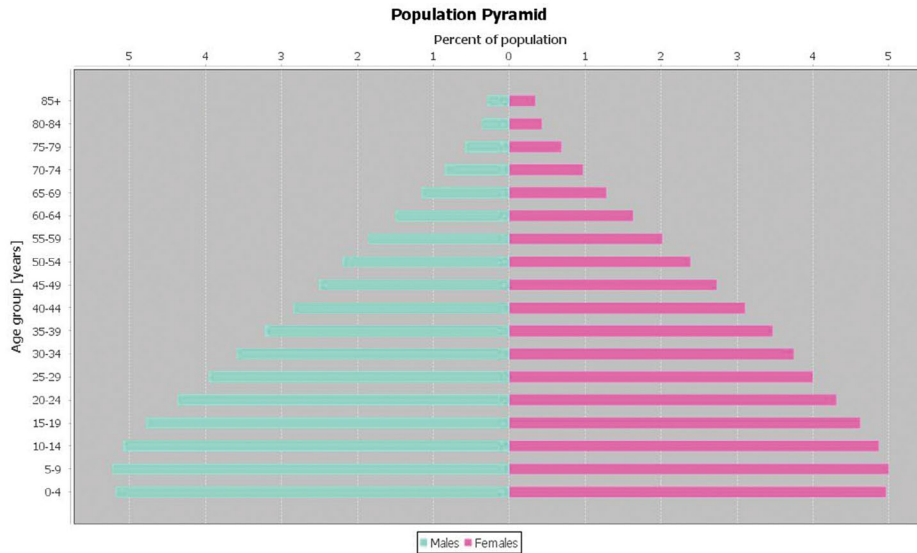


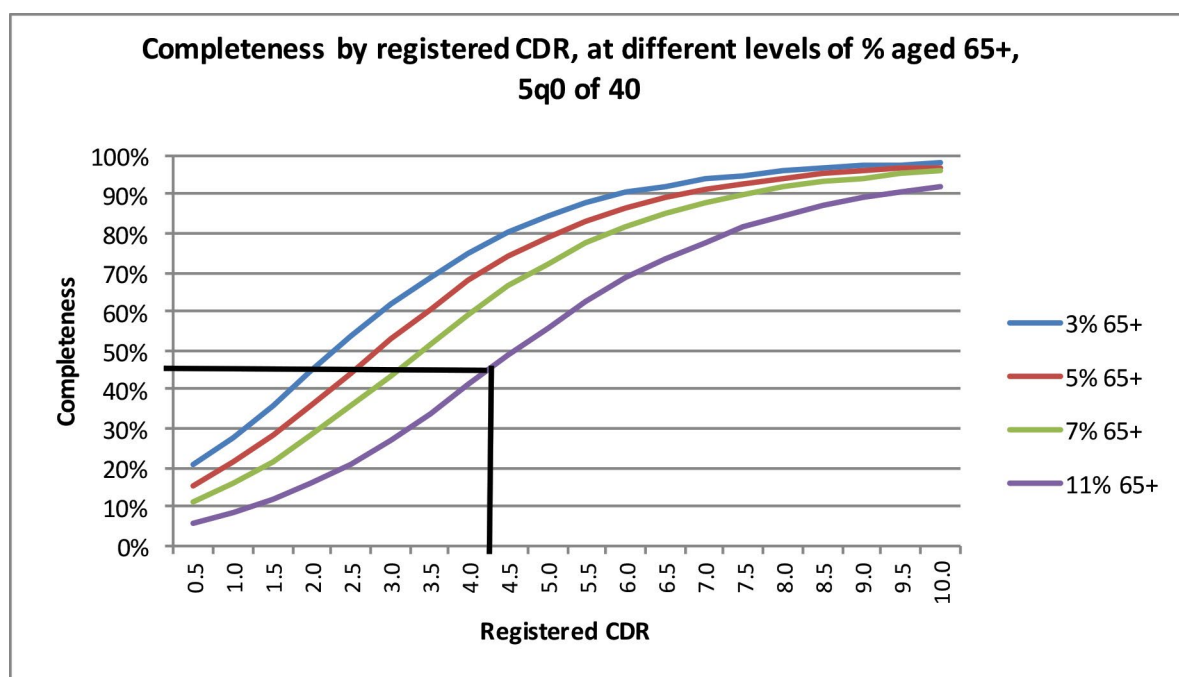
Figure 5: Example of a young population pyramid (CDR of 5 per 1000 population)



### Predicting the completeness of death reporting

The methods outlined in the previous step, where the reported CDR is compared with a comparator estimate of the CDR, require an *external* estimate of the actual number of deaths that occurred in the population to estimate the amount of undercounting. Although these estimates can be introduced via the comparator data series, there is an inherent advantage to being able to predict completeness of death registration *directly* from the input data. This is possible with a method we have developed to model the relationship between the observed CDR and that which would have been predicted given the rate at which the population is ageing and the level of child mortality (and its completeness) in the country (see Annex 1). This model allows users to directly predict their estimated completeness of death registration, given the observed CDR and the population age structure from their input data, the comparator level of child mortality, and the completeness of child mortality registration, entered into the tool in step 5 (Figure 6).

Figure 6: Modelled relationship between completeness and CDR (calculated from input data)



CDR = crude death rate

Figure 6 shows how the modelled relationship can be used to predict death registration completeness for a hypothetical country. In looking at the purple line (with 11% of the population aged 65+), and assuming the registered CDR was 4 per 1000, and a  ${}_5q_0$  of 40, the predicted level of completeness is around 45%. This precision should be sufficient to inform those responsible for improving vital registration systems about the need for action. The quantity  ${}_5q_0$  is defined as the risk of death per 1000 live births for children younger than 5. Overall, we can see the potential range in our estimate of death registration completeness given a registered CDR level and value of  ${}_5q_0$ , and how it depends on the actual proportion of the population aged 65+.

This prediction model has the advantage that it is simple to use, inherent to the input data and likely to yield levels of completeness that are close to reality. These characteristics make it particularly useful for **generating subnational estimates of completeness**. In this case, additional information on prevailing levels of  ${}_5q_0$  for different regions of the country (obtained, for example, from a previous census or survey) should be used (using the user input function in ANACONDA) for best results. The Inter-agency Group for Child Mortality Estimation (IGME)  ${}_5q_0$  country estimates can be used as a fallback. Preserving the real variation in child mortality levels and registration completeness across a country will more accurately identify areas with lower death registration completeness and thus better serve efforts to improve national death registration completeness.

### 3



#### Step 3: Age-specific mortality rates

In this step, ANACONDA checks the age and sex structure of the population, and calculates mortality rates separately for each age group and for males and females. The set of age-specific mortality rates (ASMRs) can be used to assess the quality of the mortality data by comparing the rates calculated from the input data with expected age patterns of mortality risk. By themselves, ASMRs are used to predict the expected number of deaths in a population and to check for excess or underreporting of mortality in specific age groups by comparing with ASMRs from other populations. They are therefore very relevant for the assessment of population health status. The specific relationship between age and mortality is shown in Box 3.

In contrast to the CDR, which is expressed as a rate per 1000, the ASMR is generally expressed per 100,000 because there are fewer deaths per category when allowing for age group. In countries or areas with smaller populations, the annual number of deaths at specific ages may be very small and would cause the ASMR to fluctuate unpredictably. To overcome that problem, it is advisable to calculate the ASMR over a 3- to 5-year period to average out any artificial instability due to small numbers.

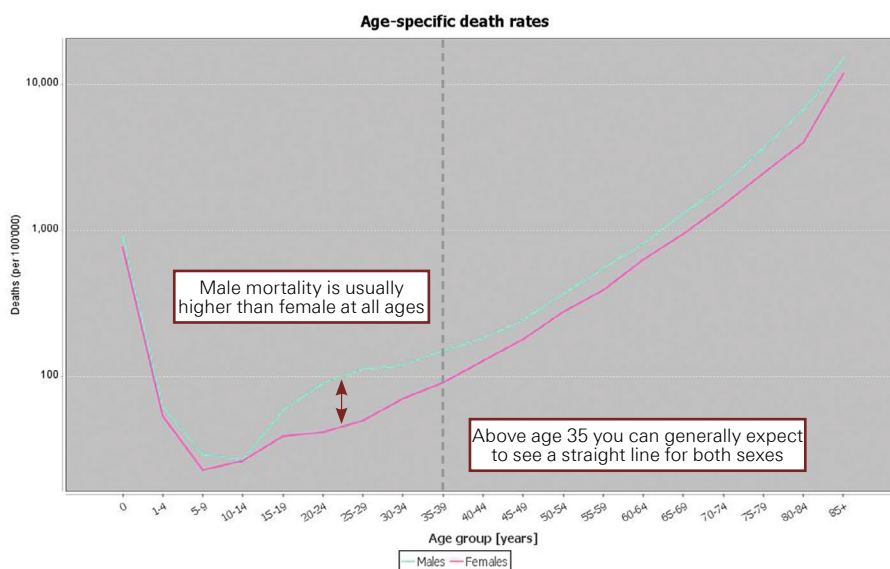


### Box 3: Age-specific mortality rates and the natural logarithm

Generally, in all settings, mortality rates are high during infancy and early childhood, and fall to their lowest levels between the ages of 5 and 14. Subsequently, mortality rates start to rise with increasing age and increase exponentially beyond age 35 or so. Therefore, the *natural logarithm (log)* of the age-specific mortality rate (ASMR) should be a straight line as age increases.

ANACONDA calculates ASMRs for each 5-year age group from the input data and illustrates them on a logarithmic graph. This allows any deviation from the expected straight line to easily be seen (Figure 7). Any significant deviations from this 'linear-in-logs' pattern are likely to reflect the underreporting of deaths at some ages compared with others, which would need to be investigated.

Figure 7: Example logarithmic graphs of ASMRs for males and females





### Understanding substep 3.1

Plotting the ASMR as shown in Figure 7 provides a simple check on whether there are any age groups where deaths are being selectively underreported. In settings where all deaths are registered, one expects to see a roughly straight line for ages 35 and older for both sexes. Because of the consistently higher male mortality, the line representing males should be higher throughout except for the very oldest age groups, where female rates sometimes are higher.

In countries with high maternal mortality in women of reproductive age and injury mortality in young adults (especially men), death rates will rise steeply around age 15 and peak before 30, after which the line will straighten out. Any departure from this linear pattern in adult death rates is likely due to deaths being selectively underreported or significant misreporting of the correct age at death.

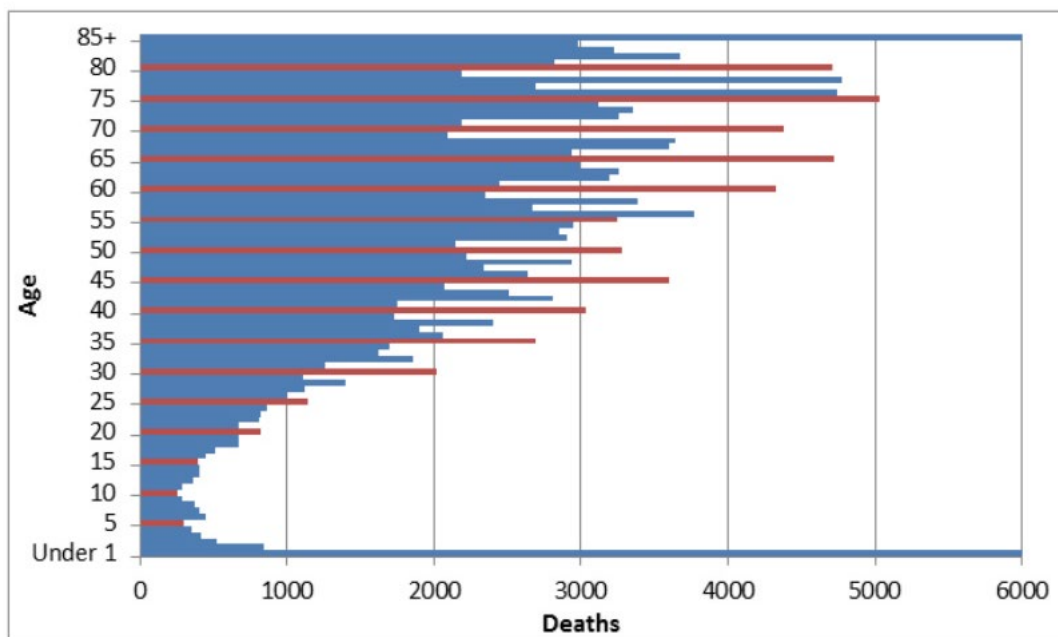
It should be noted that, in some populations severely affected by mortality from HIV/AIDS, the 'bump' in the age pattern of the log of the death rates can be quite pronounced and extend well beyond age 35 or so. Thus, when interpreting the pattern of the log of the ASMRs, it is important to know whether it is a population with high HIV/AIDS mortality, as this is likely to distort the usual pattern, typically for ages 25–55.

Although not explicitly tested for in ANACONDA, where the input data are grouped into 5-year age intervals, the quality of mortality data can be severely affected by what is known as 'age heaping', typically around ages ending in 0 or 5. In other words, families will often round up to an age ending in 0 or 5 when reporting a death to CRVS authorities. For example, rather than saying the child was 8, a parent may report they were 10. This is particularly the case for late registrations, and reporting deaths in censuses and surveys where the death took place at an earlier date. This can lead to significant distortions in the reported age pattern of deaths, and hence greatly reduce the utility of the data for understanding at what ages premature deaths are particularly high.

To assess the extent of age heaping, the distribution of deaths from the input data should be plotted according to single years of age at death, where these data are available.

An example of age heaping in the reported death data is shown in Figure 8.

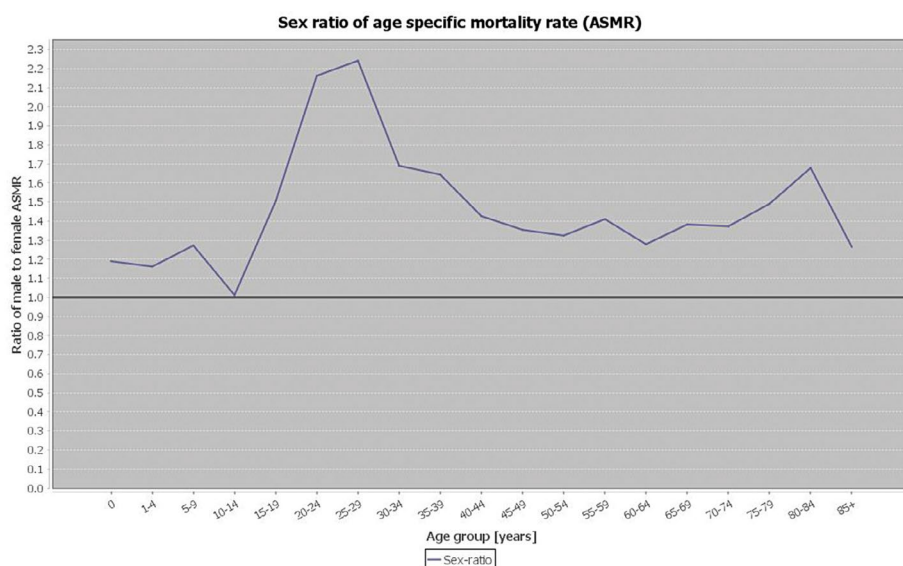
Figure 8: Example of age heaping in civil registration data



Note: Red bars show deaths for each age that ends in 0 or 5.

In addition to checking the age pattern of reported deaths, ANACONDA also assesses whether the age pattern of the comparative mortality of males and females is plausible. To better understand the age pattern of male–female differences in mortality reported in the input data, ANACONDA calculates the ratio of male to female mortality rates for each age group in **substep 3.2**. A simple and effective way of showing excess male mortality in any group is to show the ratios in a chart such as Figure 9.

Figure 9: Example of the ratio of male to female mortality rates, by age



### Understanding substep 3.2

If death rates were the same for both sexes, the ratio between the male and female death rate (ie, the male rate divided by the female rate) would be 1 for all ages. As shown in Figure 9, this is not the case. Male death rates tend to be higher than female death rates at all ages except in populations where discrimination in favour of males, and the low status of women and girls in society negatively affect their chances of survival. Typically, this excess male mortality (as measured by the ratio of death rates) will peak somewhere in the age groups 15–34 as a result of much higher male than female mortality associated with accidents, suicides and violence. In countries with high prevalence of HIV infection and and/or high maternal mortality, female mortality may exceed male mortality in some of the reproductive age groups.

A secondary, much lower peak in the ratio is often seen around 55–64 years because more males than females tend to die from chronic diseases at those ages. This is particularly pronounced in societies where males have a much higher consumption of tobacco and alcohol than females. At older ages, the sex ratio of mortality approaches 1, but the risk of death generally continues to be higher for older men than older women.

Deviations from this typical age pattern of excess male mortality are possible, depending on the country context, but should be investigated for plausibility. In particular, a higher than expected male to female mortality ratio at any age is likely indicative of differential underreporting of female deaths.

# 4

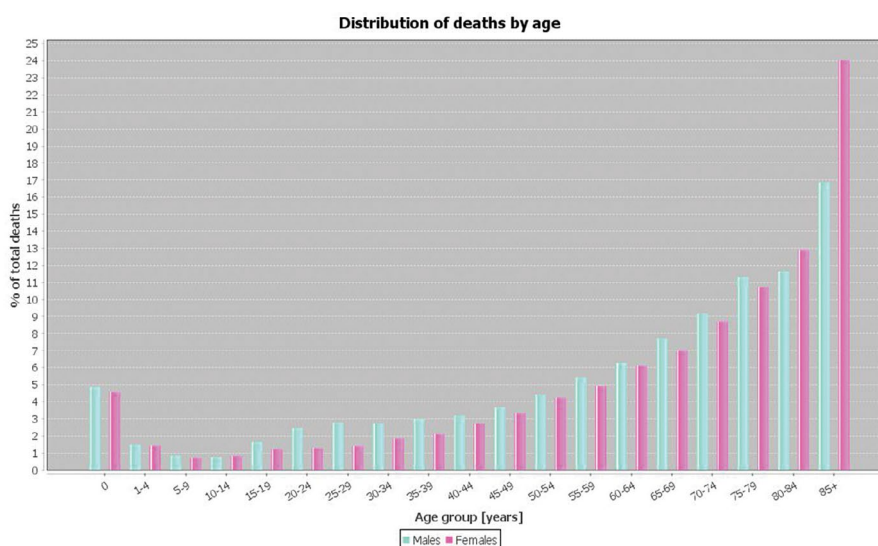


## Step 4: Age distribution of deaths

In **step 4**, ANACONDA examines the age *distribution* of registered deaths. The age distribution is the percentage of total deaths that occur at each age. The distribution would be expected to vary considerably depending on the overall level of mortality, which determines the risk of dying at each age, but also depending on the size of the population currently alive at each age. The age distribution of deaths is here shown in a histogram showing both sexes (Figure 10).

By displaying the input data in this format, users can readily spot whether there are abnormal or abrupt changes in the way that deaths are increasing with age in the input data for each sex. Presenting them together also makes it possible to show the overall shape of the sex-specific death distributions and clearly shows that higher proportions of male deaths happen at younger ages than is the case for women. Indeed, almost a quarter of all female deaths are over 85 years, which is twice the proportion for men of similar age.

Figure 10: Example age distribution of deaths for males and females



### Understanding step 4

Irrespective of the level of mortality, the heights of the bars gradually increase for each sex from the age of 5 years onwards. The higher mortality rates of males are reflected in the generally greater heights of their bars in the younger adult age groups. Hence, fewer males remain to die at the older ages. For females, the pattern is much more skewed towards the older ages, particularly the last age group (85+). This type of information is clearly of relevance to those planning or catering for the elderly population.

# 5



## Step 5: Child mortality rates

In this step, the rich and vast collection of survey- and census-derived estimates of child mortality is used to assess the plausibility of the level of child mortality calculated from the input data. The significance of this indicator is explained in Box 4. The estimates of child mortality risk calculated from censuses and surveys are likely to be more accurate than levels calculated from civil registration data since the former require ‘active’ questioning of families

about deaths of children ever born, while civil registration is a more ‘passive’ mechanism for reporting deaths that often requires families to take the initiative to register child deaths.

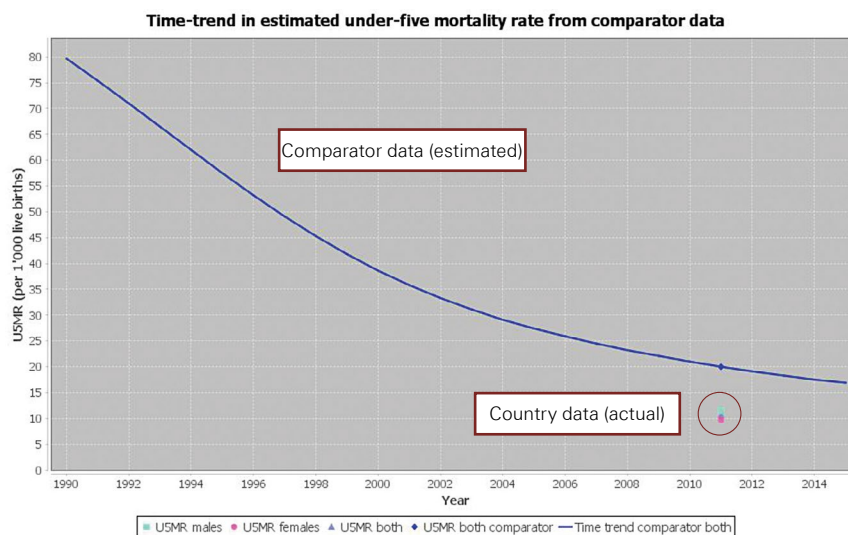
### Box 4: Child mortality measurements

More than any other age group, the mortality among children under 5 years old reflects a country’s economic, social and health conditions – hence the frequent use of child mortality as a development indicator and for health monitoring in general. This, together with the massive focus of the global health and development community over the past 5 decades on improving child survival, has led to a large amount of data in almost all developing countries, collected through censuses and surveys, on mortality levels among children under 5. In parallel, demographic research has led to a series of ‘indirect’ methods to measure the risk of child death (ie the probability of dying between birth and age 5). These methods, when applied to survey and census data, yield reliable estimates of child mortality risk.

The fact that we have relatively reliable methods to estimate true child mortality makes it useful to estimate the extent to which registration undercounts mortality of children. For instance, in country X, child mortality according to the registered deaths is 9.5 per 1000 live births. However, estimates from the Inter-agency Group for Child Mortality Estimation put it at 17.5. As a result, we can confidently say that, in country X, the registration system only registers 54% of all child deaths, or undercounts child deaths by 46%.

In **step 5.1**, ANACONDA calculates the 5q0 from the input data using life table techniques. This risk, calculated from the civil registration data, can then be directly compared with the risk of child death calculated from censuses and surveys, or, as is more commonly done, with some trend line fitted to the series of estimates derived from different sources (Figure 11). The reliability of the under-5 rate is dependent on both the accuracy and completeness of death and birth reporting as the denominator is ‘number of live births’.

Figure 11: Comparison of input data and estimated under-5 mortality rates



USMR = under-5 mortality rate



## Understanding substep 5.1

If there are indications from the CDR (step 2) that deaths are underreported, it is most likely to occur for young children, especially if deaths happen very soon after birth (during the neonatal period from birth to 28 completed days after birth). Indeed, it is not uncommon that many deaths during the early neonatal period are misclassified as stillbirths and therefore neither the birth nor the death of the child is recorded. This is very poor practice, as data on both events are critical to inform strategies to improve maternal and child health services.

To check whether child deaths are being underreported in the input data, ANACONDA compares the under-5 mortality rate (also referred to as 5q0) calculated from the input data with the corresponding level, estimated by the IGME, based on fitting a line to the universe of (plausible) child mortality estimates derived from censuses and surveys.

In Figure 11, which displays data from two different sources, the extent of underreporting of child deaths by the civil registration is measured by the *relative* difference between the calculated risk from the input data and the comparator data (blue line), rather than the absolute difference from the estimated trend line.

In the above case, the relative difference is calculated as  $(22.5 - 13.1) \div 22.5 \times 100$ , which equals 42%, suggesting that 4 out of 10 children who die are not being registered.

ANACONDA has to construct a life table to be able to calculate the risk of dying for children, as opposed to simply dividing under-5 deaths by live births (since the comparable metric from the IGME is the *risk* of child death). Therefore, the life table calculated from the input data on deaths and population alive at different ages has been included in **substep 5.2**. The life table provides other measures of direct public health importance, such as the probability of dying at any age for a particular year (Box 5). However, if serious underreporting of deaths (and births) is suspected, the life table will also be biased and will substantially overestimate indicators such as life expectancy at various ages, and hence should be used and interpreted with great caution.

### BOX 5: The life table

A life table shows, for each age, what the probability is that a person will die before his or her next birthday (probability of death). We do this by assuming that the mortality rates observed in a given year (ie the period) at a given age (x) will stay the same throughout a person's lifetime. ANACONDA calculates, based on the input mortality and population data, a period life table for that given year. The life table applies these period mortality rates, at each age, to an artificial or imaginary population of 100,000 births, and follows them throughout their (imaginary) life by applying the fixed or period death rates to this cohort of 100,000 births until they all die out. In reality, it may take 100 or more years for a given birth cohort to die out, but we can model their progress through life instantaneously by applying the period mortality rates in a given year and assuming they apply throughout the life of the cohort.

#### Example of an abridged life table

Age	Years in interval	Deaths	Probability of dying	Probability of surviving	Individuals surviving	Deaths in interval x	Years Lived in interval x	Cumulative years lived	Expectancy of life at age x
x	n	${}_nD_x$	${}_nO_x$	${}_nP_x$	$l_x$	${}_n d_x$	${}_n L_x$	$T_x$	$e_x$
0	1	4608	0.012609036	0.987390964	100000	1260.903615	98865.18675	75.14971215	75.14971215
1-4	4	876	0.002198523	0.997801477	98739.09639	217.0801595	394435.3932	74.16106.028	75.10810104
5-9	5	515	0.001246196	0.998753804	98522.01623	122.7777798	492303.1367	70.21670.635	71.27006637
10-14	5	690	0.001655689	0.998344311	98399.23845	162.9185102	491588.896	65.29367.498	66.35587431
15-19	5	3412	0.007969989	0.992030011	98236.31994	782.9423639	489224.2438	60.37778.602	61.4617751
20-24	5	4989	0.011686728	0.988313272	97453.37757	1138.911138	484419.61	55.48554.359	56.93547517
25-29	5	4786	0.011707637	0.988292363	96314.46643	1127.614799	478753.2952	50.64134.749	52.57917046
30-34	5	4230	0.011312732	0.988687268	95186.85164	1076.823295	473242.1999	45.85381.453	48.17242481
35-39	5	3558	0.010975418	0.989024582	94110.02834	1032.89689	467967.8995	41.12139.253	43.69501663
40-44	5	3739	0.011912594	0.988087406	93077.13145	1108.790101	462613.682	36.44171.354	39.15216657
45-49	5	4189	0.01420508	0.98579492	91968.34135	1306.417658	456575.6626	31.81557.672	34.59405297
50-54	5	5057	0.020663715	0.979336285	90661.92369	1873.412161	448626.0881	27.24982.009	30.05652096
55-59	5	6498	0.032645228	0.967354772	88788.51153	2898.521208	436696.2546	22.76355.921	25.63795566
60-64	5	7405	0.048748375	0.951251625	85889.99032	4186.997438	418982.458	18.39659.667	21.41879001
65-69	5	8301	0.079397038	0.920602962	81702.99288	6486.975614	392297.5254	14.20677.209	17.38831294
70-74	5	10525	0.146053773	0.853946227	75216.01727	10985.58309	348616.1286	10.28379.683	13.6723496
75-79	5	11658	0.231176506	0.768823494	64230.43418	14848.56732	284030.7526	6.79763.5547	10.58320037
80-84	5	12139	0.369308934	0.630691066	49381.86685	18237.1646	201316.4228	3.95732.8021	8.013727049
85+	5	15539	1	0	31144.70225	31144.70225	194416.3793	194416.3793	6.242357938

In the example, the mortality rates from the input data are first converted into probabilities of dying at each age  $x$ , conventionally designated as  $q_x$ , for each year of age  $x$ . These probabilities are shown in the third column, whereas the fourth column shows the probability of surviving for the cohort. The fifth column,  $l_x$ , shows the number of survivors or the (imaginary) life table population who reach a given year of age. For example, of the 100,000 of the life table population who had just been born (ie were exact age 0), 825 could be expected to die before reaching their first birthday, based on the period mortality rates observed in the population from which the input data came. Similarly, of the 99,175 who reached their first birthday (age 1), 62 would be expected to die between their first and fifth birthdays. This procedure is continued to the end of the life table - in this case, 85+ when all 100,000 will have died. From this, we compute the number of deaths (column  $d_x$ ) at each age among the imaginary cohort of 100,000 births that we started off with.

The  $L_x$  column denotes the total number of years of life lived by survivors at age  $x$  before they reach age  $x + 1$ . If no-one died between age  $x$  and  $x + 1$ ,  $L_x$  would simply be equal to  $l_x$ . But since some will die ( $d_x$ ) at age  $x$ , we assume they die, on average, midway between ages  $x$  and  $x + 1$ . In other words, we assume that each of the people dying after reaching their  $x^{\text{th}}$  birthday lived only one-half year after their  $x^{\text{th}}$  birthday. Accordingly, among the number of survivors,  $l_x$ , who attain age  $x$ , the years of life lived by those dying during that year of age is  $\frac{1}{2}d_x$ . The years of life lived by the survivors is  $l_x + 1$ , which is equal to  $l_x - d_x$ . The sum of  $\frac{1}{2}d_x$  and  $l_x - d_x$  is the total number of years lived within that year of age.

The column headed  $T_x$  is the total number of years of life lived by the life table cohort beyond age  $x$  until the end of the life table, when all of the original 100,000 births will have died, again assuming that each of the persons dying lived one-half year after the last birthday.

Finally, to compute the *expectation of life* ( $e_x$ ), at birth or at any age, just divide the total number of years lived by the cohort ( $T_x$ ) by the number of those who survived to that exact age ( $l_x$ ) or:

- $e_x = \text{average years of life remaining for a person at age } x = T_x / l_x$ .

Life expectancy at birth,  $e_0$ , is the most commonly used life table measure, since it summarises period mortality rates over the entire lifetime of a population. Other commonly used measures extracted from a lifetable include:

- ${}_5q_0$  – the probability that a newborn at exact age 0 will die before his or her 5<sup>th</sup> birthday

${}_{45}q_{15}$  – the probability that an adult who reaches his or her 15<sup>th</sup> birthday will die before he (or she) reaches age 60, 45 years later.

Always remember that a life table, such as the one above, assumes that the mortality rates observed in a country or subpopulation during the specified period will continue unchanged over time. In that sense, the life table is a summary of the implications of current mortality rates if these were applied to an imaginary birth cohort of 100,000 (called the ‘radix’ of the life table). Life expectancy is an easy-to-understand indicator of mortality levels in a particular population at a particular time, and it is frequently used to compare the overall mortality levels of various populations.

### (3) Cause of death analysis

6



#### Step 6: Classification of deaths into broad cause groups

In addition to the mortality analysis explained above in the first five steps, ANACONDA provides a detailed framework for assessing the plausibility and quality of COD data (Box 6). This analysis is important because public health interventions target specific diseases or injuries, as do health policies and some health services, and thus require accurate and timely data on mortality due to those diseases. For example, the public health strategies to control

lung cancer are very different from those developed to control cervical cancer. Hence the need for accurate and timely cause-specific mortality data.

### Box 6: Standard grouping of diseases into three major groups

As a first step, it is important to check the data to ensure that broad categories of causes of death (COD) are reliably being reported. For example, is the fraction of deaths due to injuries reported in the input data about what one would expect, irrespective of the specific causes of those injury deaths (eg suicide, homicide, accident). Once this has been assessed in **step 6**, subsequent steps will interrogate the data in more cause-specific detail.

The thousands of possible CODs contained in ICD 10 were classified into three very broad groups that are useful for health policy debate (Murray and Lopez 1996):

1. Group I. Infectious and parasitic diseases (eg tuberculosis, pneumonia, diarrhoea, malaria, measles); maternal and neonatal causes (eg, maternal haemorrhage, birth trauma); malnutrition.
2. Group II. Noncommunicable diseases (eg cancer, diabetes, heart disease, stroke); mental health conditions (eg schizophrenia).
3. Group III. Injuries (eg accidents, homicide, suicide).

The expected percentage distribution of CODs across these three groups varies in different countries according to where the country stands in relation to the 'health epidemiological transition – an interrelated set of changes in demographic structures, patterns of disease and risk factors. Population ageing can be observed in all populations, and results from falling fertility rates and declines in mortality rates, initially among children under 5 years old, and then among middle and older age adults. This results in epidemiological changes, including a shift in the main COD structure away from infectious diseases, such as gastrointestinal diseases and pneumonia (diseases traditionally associated with poorer countries), towards noncommunicable diseases such as cardiovascular disease, stroke and cancers, which mainly affect older adults. As countries develop, risk factors for infectious diseases (eg undernutrition, unsafe water and poor sanitation) are brought under control, while risk factors for chronic diseases (such as being overweight, using alcohol and tobacco) increase. But, in practice, the transition does not always happen quite so cleanly, with many countries or parts of countries facing a dual burden of disease associated with both group I and group II conditions (eg South Africa).

A simple but effective way of checking the plausibility of mortality data is to compare the observed patterns of CODs with what would be expected given the probable levels of life expectancy. Generally, countries with low life expectancy are characterised by high levels of mortality due to infectious and parasitic diseases, especially in childhood, along with high maternal mortality (group I causes). As life expectancy rises, the pattern of mortality changes, with more deaths occurring in older age groups due to noncommunicable conditions such as cardiovascular diseases and cancers (group II causes). The proportion of deaths due to injuries typically remains fairly constant as life expectancy increases (Table 1).

The model-based percentages in Table 1 show how the percentage of deaths assigned to various causes in each of the groups is expected to change as life expectancy increases. Thus, a country with an average life expectancy of 55 years would typically have about 22% of deaths due to group I causes and 65% due to group II causes. A country with a life expectancy of 65 years would typically have a smaller percentage of deaths due to group I conditions (around 13%) and correspondingly more deaths due to group II conditions (74%). As these are model-based percentages derived from the WHO database, it is unlikely that any country would fit exactly these proportions, but significant departures from these broad patterns would suggest potential problems with the certification or coding of CODs.

In **step 6**, ANACONDA distributes all the usable ICD-coded deaths in the input dataset into the three major groups outlined in Box 6 and taken from the 1990 GBD study. It also shows the fraction of deaths assigned to what are termed 'unusable' COD codes. This is illustrated in Figure 12.

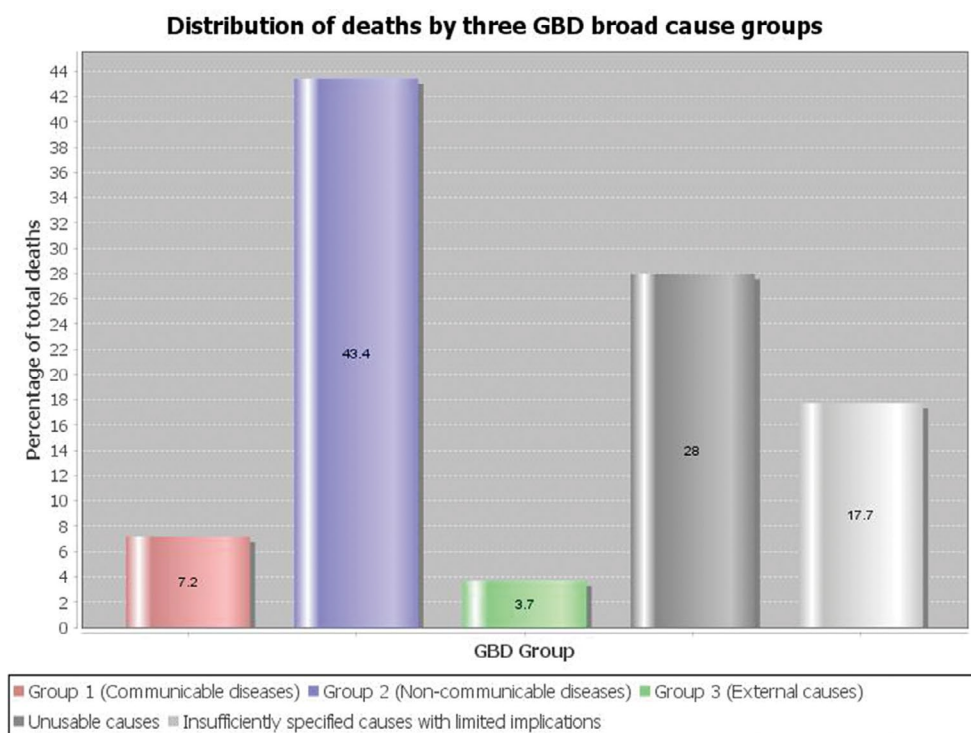
**Table 1: Expected distribution of COD according to life expectancy, by broad groups**

Group	Life expectancy (years)			
	55	60	65	70
I	22%	16%	13%	11%
II	65%	70%	74%	78%
III	13%	14%	13%	11%

Source: AbouZahr et al (2010)



Figure 12: Distribution of deaths by three major disease groups



GBD = global burden of disease



### Understanding step 6

Nosologists<sup>4</sup> have identified a number of causes in ICD 10 that cannot or should not be used as the underlying cause of death (UCOD) (eg ‘heart failure’), or that contain no information about the probable UCOD (eg ‘ill-defined causes’). Further, the International Classification of Diseases contains a number of codes that, while broadly useful for public health purposes (such as ‘ill-defined sites of cancer’), are not specific enough to guide public policy dialogue. These are also included among the unusable codes in certain cases.

A common problem in many countries is the large number of deaths that are not properly certified and hence end up with some unusable code. Showing the three broad disease groups together with this category of vague and poorly defined deaths is an effective way to summarise the extent of the challenges facing doctors, and coders and those responsible for health information systems in the country. If the proportion of unusable codes is large (more than 10%), it will bias the distribution of deaths shown in the three major groups (since it is unlikely that unusable codes are equally or proportionally distributed across the three broad categories) and the data will not represent the true health status of the population.

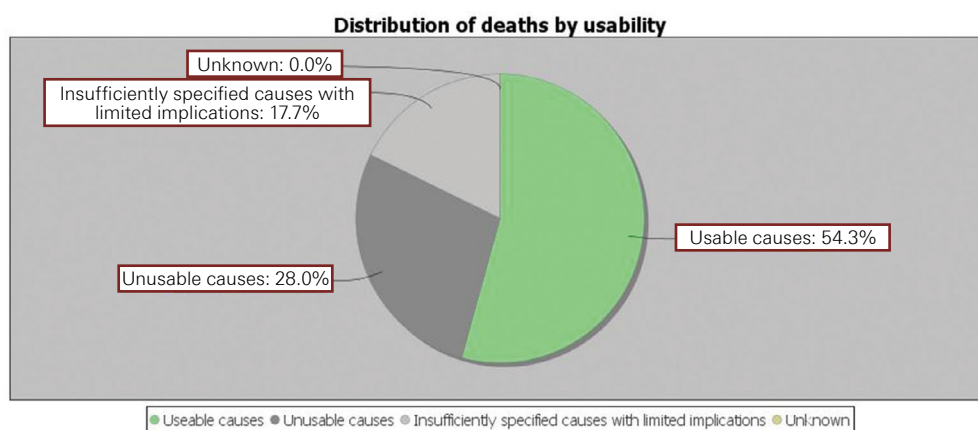
The bias introduced into the COD distribution can be more serious if the type of unusable cause given is one where even the broad category of the underlying cause cannot be correctly determined. For example, when septicaemia is written on a death certificate, the death may be the result of communicable disease, noncommunicable diseases, or accident and injury. This is because septicaemia is the *immediate* COD, but was caused by some other condition such as an accident, cancer or an infectious disease (Figure 12).

<sup>4</sup> Nosology is a branch of medicine that deals with the classification of diseases.

As countries develop, and communicable diseases, malnutrition, and maternal and neonatal conditions are brought under control, more and more deaths will occur from noncommunicable diseases among the growing number of adults and the elderly. A simple way to check if this epidemiological transition is being reliably reflected in the mortality input data is to calculate the ratio of deaths in group II to group I. This ratio is a powerful descriptor of how fast disease patterns are changing in the country. The higher the ratio, the further the country has moved through this evolution. ANACONDA computes this ratio in step 6 and compares it with the group II/group I ratio calculated from the comparator (GBD region in which the country is situated – see Figure 1). If the two ratios differ substantially, the input data should be carefully reviewed in case one or other of the two broad disease categories (group I or group II) is being differentially underdiagnosed.

To further illustrate the evidence base provided by the input dataset, all the unusable and insufficiently specified codes found in the dataset are extracted and shown in a pie chart (Figure 13). The figure shows that any policy based on the dataset below will have made use of only 55% of the available data. With such a high fraction of unusable codes, the whole dataset is of no use to policy and can seriously mislead users about the real health problems of the population.

Figure 13: Distribution of deaths by usability



## 7



### Step 7: Quality of cause of death data

The most common reason for poor quality COD data is that doctors do not know how to fill out the death certificate correctly, particularly the section allowing the selection of the underlying UCOD. The various errors introduced by doctors when certifying the COD have led to what are collectively known as ‘ill-defined’, ‘garbage’ or ‘unusable’ codes, as discussed

above. **Step 7** investigates, in detail, the contents of the unusable codes and other quality aspects of the COD input data. The step is broken down into eight substeps.

**Step 7.1** introduces the International Form of Medical Certificate of Cause of Death (World Health Organization 2016), the use of which is recommended for improving the quality of COD data.<sup>5</sup>

Although ANACONDA does not specifically analyse the form used to collect the input data, users are encouraged to check how closely the form used in their dataset resembles the international standard. The design of the international form encourages the certifying physician to report the sequence of events leading to death, and hence facilitates the selection of the UCOD by correct application of the ICD 10 coding rules by trained coders. For public health, it is the UCOD that is important, rather than the immediate cause that led to death (such as septicaemia), as this is the disease or condition that initiated the train of morbid events that led to death, and which public health efforts should seek to prevent. All tabulations of COD statistics should be based on the UCOD and not on the immediate cause.

In **substep 7.2**, ANACONDA compiles all deaths in the input data according to the 22 chapters of the ICD, and shows the number of unusable codes within each chapter (Table 2). By indicating the fraction of deaths belonging to each chapter and the fraction of unusable codes attributed to each chapter, it is possible to immediately see where the unusable causes are coming from and which are the major areas of concern for incorrect COD certification. Are there disease chapters that contribute more than would be expected from their share of deaths? For example, if an ICD chapter accounts for only 2% of deaths but contributes 10% of all unusable codes, then certification practices for diseases contained in this chapter require special attention. This type of analysis makes it possible to better target intervention efforts and gain important insights into likely biases in the COD patterns suggested by the data.

**Table 2: Example ANACONDA output: total deaths by ICD chapter, with percentage unusable and insufficiently specified codes in each**

ICD chapter	Description	ICD code range	Total deaths	% of total deaths	Total unusable causes	Total unusable level 1/2/3	Total unusable level 4	% of total unusable causes
1	Chapter I: Certain infectious and parasitic diseases	A00-B99	39,405	7.3	7,490	7,383	107	3.7
2	Chapter II: Neoplasms	C00-D48	57,823	10.8	6,292	6,292	0	3.1
3	Chapter III: Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	D50-D89	4,090	0.8	2,770	2,770	0	1.4
4	Chapter IV: Endocrine, nutritional and metabolic diseases	E00-E90	39,361	7.3	4,013	4,013	0	2.0
5	Chapter V: Mental and behavioural disorders	F00-F99	791	0.1	357	357	0	0.2
6	Chapter VI: Diseases of the nervous system	G00-G99	7,525	1.4	3,599	2,655	944	1.8
7	Chapter VII: Diseases of the eye and adnexa	H00-H59	10	0.0	10	10	0	0.0
8	Chapter VIII: Diseases of the ear and mastoid process	H60-H95	35	0.0	31	31	0	0.0
9	Chapter IX: Diseases of the circulatory system	I00-I99	187,759	35.0	73,959	42,424	31,535	36.7
10	Chapter X: Diseases of the respiratory system	J00-J99	82,607	15.4	58,089	5,757	52,332	28.9
11	Chapter XI: Diseases of the digestive system	K00-K93	21,187	3.9	3,503	3,503	0	1.7
12	Chapter XII: Diseases of the skin and subcutaneous tissue	L00-L99	2,193	0.4	68	68	0	0.0
13	Chapter XIII: Diseases of the musculoskeletal system and connective tissue	M00-M99	1,820	0.3	967	967	0	0.5
14	Chapter XIV: Diseases of the genitourinary system	N00-N99	20,539	3.8	5,141	5,141	0	2.6
15	Chapter XV: Pregnancy, childbirth and the puerperium	O00-O99	1,570	0.3	3	3	0	0.0
16	Chapter XVI: Certain conditions originating in the perinatal period	P00-P96	9,867	1.8	712	0	712	0.4
17	Chapter XVII: Congenital malformations, deformations and chromosomal abnormalities	Q00-Q99	4,297	0.8	0	0	0	0.0
18	Chapter XVIII: Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	R00-R99	20,673	3.8	20,254	20,254	0	10.1
19	Chapter XIX: Injury, poisoning and certain other consequences of external causes	S00-T98	0	0.0	0	0	0	0.0
20	Chapter XX: External causes of morbidity and mortality	V01-Y98	35,447	6.6	14,017	5,702	8,315	7.0
21	Chapter XXI: Factors influencing health status and contact with health services	Z00-Z99	0	0.0	0	0	0	0.0
22	Chapter XXII: Codes for special purposes	U00-U85	0	0.0	0	0	0	0.0
Total			536,999	100.0	201,275	107,330	93,945	100.0



### Understanding substep 7.2

Unusable codes can be found in most chapters, and knowing the proportion in each chapter is a useful first step towards identifying poor certification/coding practices associated with certain diseases, particularly if the proportion of unusable codes is far greater than expected. Chapter 18 of the ICD ('Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified') contains the vast majority of ill-defined conditions for which no specific diagnosis can be made.

One would expect chapter 18 to account for a high proportion of ill-defined and unusable codes. However, in general, the highest proportions are found in chapter 9 ('Circulatory system'), chapter 10 ('Respiratory system') and chapter 2 ('Neoplasms'), suggesting that certification and coding improvement strategies should focus on the major diseases contained in these chapters.

For many countries, particularly those that have been successful in registering most deaths, the major challenge to improving the utility of CRVS data lies in improving the accuracy of COD data. This in turn poses two challenges – ensuring that the UCODs that are *not* considered as unusable are diagnosed correctly and reducing the fraction of unusable codes in the data. A recent meta-analysis of studies to evaluate diagnostic accuracy of the UCODs occurring in hospitals found systematic errors, resulting in an incorrect diagnosis in 25–65% of cases (Rampatige et al 2014). This type of diagnostic evaluation, although important, requires special research studies on a sample of deaths in a sample of hospitals (Rampatige et al 2014), and is outside the scope of ANACONDA.

### Classifying ‘unusable’ cause of death codes according to type

In **substep 7.3**, ANACONDA breaks up the unusable causes and classifies these into five different categories, according to their type, as follows:

1. *Category 1*: Symptoms, signs and ill-defined conditions mostly drawn from R00–R99 in ICD 10
2. *Category 2*: Impossible as UCODs. Includes causes such as ‘essential hypertension’ and ‘atherosclerosis’
3. *Category 3*: Intermediate CODs. Causes that have been precipitated by the underlying cause
4. *Category 4*: Immediate CODs, such as heart or respiratory failure
5. *Category 5*: Insufficiently specified causes within ICD chapters within a larger disease category. These include causes such as ‘ill-defined site of cancer’ and ‘ill-defined injuries’. Use of these codes is unhelpful in guiding prevention efforts. Such efforts are usually cause specific (eg lung cancer, early diagnosis of breast and prostate cancer).

The classification provides detailed insight into the type and frequency of unusable codes in the country input data. It is helpful to determine the most effective intervention necessary for correcting the data and for monitoring the impact of the interventions over time. Countries are therefore urged to carefully review the patterns of unusable codes in their data and design certification improvement strategies that are aligned with the error patterns in their unusable codes.<sup>6</sup> The first category in the classification corresponds to chapter 18 of the ICD and amalgamates symptoms, signs and ill-defined conditions that should never be assigned as the UCOD, or at least limited to the few times when it really is not possible to assign a more specific COD, such as sudden deaths that could not be autopsied (R96–99).

The next three categories group together conditions that cannot, or should not, be used as UCODs, because they are nosologically impossible, or because they represent the intermediate or final condition before death.

The fifth category has also been included among the unusable codes because more specific diagnoses could have been made. This category includes causes such as B99 ‘Other and unspecified infectious diseases’, C80 ‘Malignant neoplasms without specific site’, F50 ‘Eating disorders’ and I99 ‘Other unspecified disorders of circulatory system’. They may help define the amount of mortality within a broad group (eg, cancer or injuries), but their lack of specificity restricts their value for planning and evaluating public health programs.

In the dataset described in Table 3, 1 in 3 deaths was given an unusable COD of some sort, with most of them being intermediary causes and insufficiently specified causes.

**Table 3: Classification of unusable causes by type**

Unusable codes classification	No. of deaths with unusable codes	% of total causes	% of total unusable and insufficiently specified causes
Category 1: ‘Symptoms, signs and ill-defined conditions’	18,004	3.4	8.9
Category 2: ‘Impossible as underlying causes of death’	4,969	0.9	2.5
Category 3: ‘Intermediate causes of death’	47,188	8.8	23.4
Category 4: ‘Immediate causes of death’	11,496	2.1	5.7
Category 5: ‘Insufficiently specified causes within ICD chapters’	119,618	22.3	59.4
Total unusable and insufficiently specified causes	201,275	37.5	100.0

## Classifying 'unusable' cause of death codes according to policy implications

In **substep 7.4**, an alternative approach to classifying the *same universe* of unusable and insufficiently specified codes is proposed. It regroups unusable causes according to their potential impact for guiding or misguiding public policy to prevent premature deaths. In this classification, proposed by an international expert group,<sup>7</sup> four levels of ICD codes that should be avoided are defined, depending on how serious their impact is for misguiding public policy.

- **Level 1 – Codes with serious implications likely to have a very high impact** for health policy. These are causes for which the true UCOD could belong to more than one broad GBD group, such that it cannot even be determined whether the true cause was a communicable disease, a noncommunicable disease or the result of an injury. These are clearly the most serious of the unusable codes, since they could potentially alter our understanding of the broad pattern of CODs in the population.
- **Level 2 – Codes with substantial implications likely to have a high impact.** These are causes for which the true COD is likely to belong to only one of the three broad GBD groups. In other words, the broad COD group to which the unusable cause belongs can be determined. These unusable causes are less serious than level 1, since they do not alter our understanding of the broad composition of CODs in the population. They do, however, affect our knowledge of leading CODs.
- **Level 3 – Codes with important implications likely to have medium impact.** These are causes for which the true UCOD is known to be within the same ICD chapter. For instance, a death assigned to 'ill-defined site of cancer' indicates that the true COD was cancer, but does not specify the site. Knowledge about the site of cancer is important for public health policy, since different strategies are applied for different types (sites) of cancer.
- **Level 4 – Codes with limited implications likely to have low impact.** These are diagnoses for which the true COD is likely to be confined to a single disease or injury category. For example, 'unspecified stroke' would still be assigned as a stroke death, and not to some other disease category. The implications for public policy of unusable causes classified at level 4 will therefore generally be minor. Hence, it is suggested that countries focus energies on eliminating the three other levels. The fourth level is included for those countries that have fairly accurate COD data, but could do better because of the availability of diagnostic tests.

Including the classification of unusable codes in ANACONDA allows countries to see the comparative importance of the four levels of unusable codes. It would then be up to the country to decide the level it would be interested in investigating further. For most countries, the important levels to work on would be levels 1, 2 and 3, which would contain the most important unusable codes – those likely to have substantial implications for correctly understanding disease and injury patterns in the population. Level 4 is likely to contain those codes that, although not as specific as is allowed for in the ICD, would demand considerable diagnostic sophistication and equipment to correct, and hence might not be possible to resolve in all countries and circumstances. Besides, as stated above, their impact for policy would be minor.

**Table 4: Distribution of unusable codes according to severity levels**

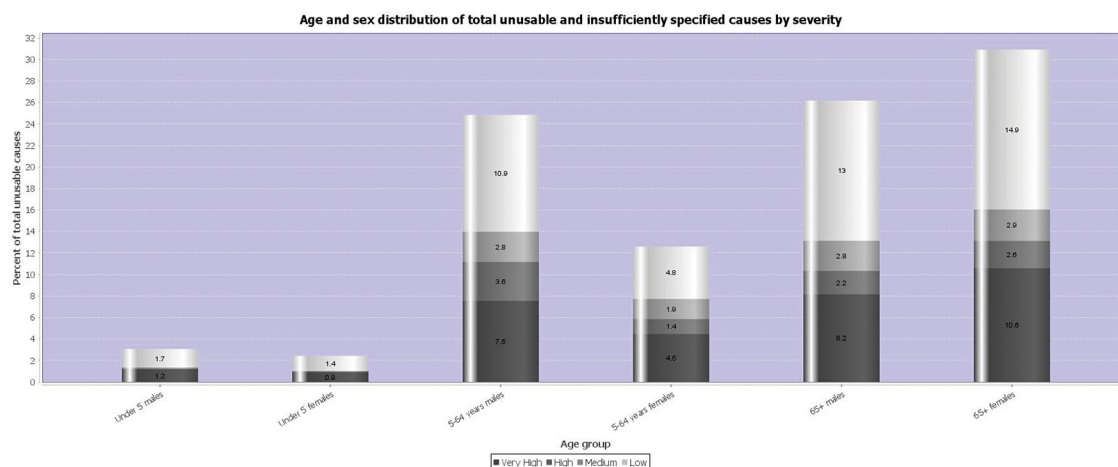
Unusable codes severity	No. of deaths with unusable codes	% of total causes	% of total unusable and insufficiently specified causes
Very high	66,226	12.3	32.9
High	19,845	3.7	9.9
Medium	21,259	4.0	10.6
Low	93,945	17.5	46.7
Total unusable and insufficiently specified causes	201,275	37.5	100.0

<sup>7</sup> Expert Group meeting on garbage codes, Melbourne School of Population and Global Health, 27–28 March 2017.

In the dataset examined in Table 4, approximately 30% of all deaths have been assigned an uninformative cause. However, close to 45% of these have a very high impact on the COD distribution, as the true underlying cause could be in a different broad group of diseases or injuries, 22% are within the same broad group and 17% within the same ICD chapter. The 16% that are likely to be within the same specific disease group, such as 'stroke unspecified', can probably be ignored, as they would not distort the diseases pattern for public health purposes. This additional information has only minor importance for overall disease and injury control strategies in a country, which is not the case for levels 1, 2 or 3. In such cases, as in the example in Table 4, it would seem advisable when formulating a strategy to reduce unusable causes to focus on levels 1–3, with priority given to reducing the use of unusable codes classified as level 1.

To understand further whether unusable codes are more frequent in certain population groups, we examine the age and sex distribution of the four levels identified. As people age, they are more likely to suffer from diseases and conditions concurrently. This makes it difficult for doctors to identify a single underlying cause or even a specific cause. As a result, it is not surprising to see that unusable CODs increase with age, as shown in Figure 14. The fact that in most countries women survive to a higher age than men is likely to be the explanation for the taller bar for women in the 65+ age group.

Figure 14: Age and sex distribution of unusable causes

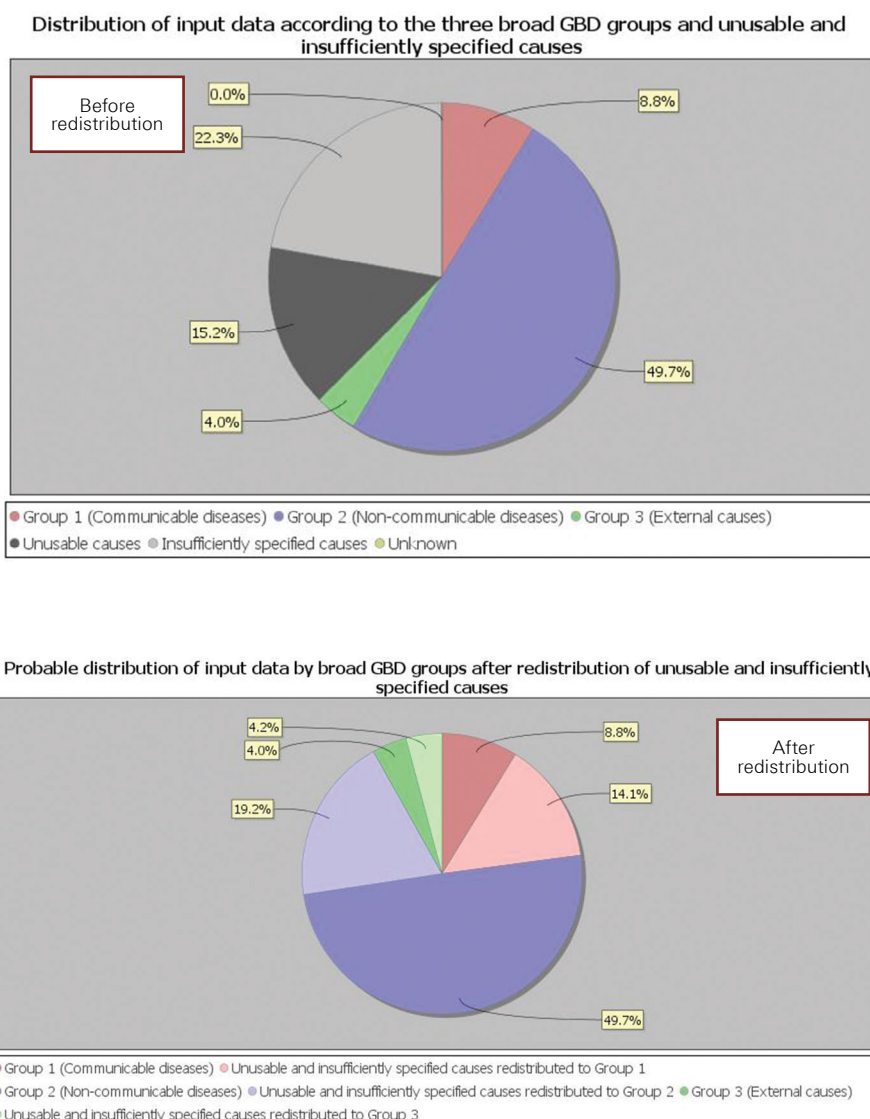


### Understanding substep 7.5

Figure 14 clearly shows that CODs classified as unusable are much less common for children than for any other age group. Deaths that occur above age 65 years have the highest proportion of all four types of unusable codes. As people age, they are likely to suffer from a number of diseases and conditions concurrently. This makes it difficult to identify a single UCOD as required by the ICD rules. As a result, it is commonly observed that the proportion of unusable ICD mortality codes is higher in the oldest age groups, as shown in Figure 14. In addition, because many deaths in older people occur outside hospitals, the certifying doctor may not have access to the decedent's medical records and may assign a vague, unspecific COD based on information provided by the family.

The purpose of the visualisations in **substep 7.6** is to show the significant impact that unusable codes can have on distorting the likely true COD composition in the population, even when the data are compiled into major disease groups. This is done by showing the death distribution in a pie chart before and after the redistribution of unusable codes (Figure 15). The redistribution algorithm used in the second pie chart to assign unusable codes to their most likely broad cause group is identical to that used for the GBD study, and is based on clinical observations and opinion about probable misdiagnoses, statistical algorithms and proportional redistribution according to reported COD patterns (Lozano et al 2013).

Figure 15: Example ANACONDA output: distribution of data according to three broad GBD groups, before and after redistribution



### Understanding substep 7.6

As clearly demonstrated in the two pie charts in Figure 15, when the unusable fraction of all deaths is high (dark grey segment in the first chart), the implications for understanding the most likely COD pattern in the population can be substantial. This is evident from the second pie chart, where the unusable CODs have been redistributed according to complex algorithms based on the GBD, with each disease group reallocated to the three broad categories. Most vague and ill-defined diagnoses of death are likely to be redistributed to group II (noncommunicable diseases), with lesser amounts going to communicable diseases and injuries. The proportionate impact of redistributing unusable codes, however, is often greater for group I than group II (compare the relative sizes of the two red slices versus the relative sizes of the two blue slices in the two pie charts in Figure 15) and sometimes the injury slice gains the most, relatively speaking.

Countries with different age structures, socioeconomic development and medical practices will show different patterns of unusable codes. In **substep 7.7**, we return to the classification of unusable codes into severity levels, and ask: In each of the four levels identified in step 7.4, which were the most commonly misused CODs? To provide further insight into the distribution of causes at each level, ANACONDA offers the possibility of drilling down into the different levels of unusable codes to discover which disease groups they contain. In other words, users will be able to identify the specific set of diagnostic practices that are responsible for the vast majority of unusable codes within each level.

Given the large number of unusable CODs classified into each of the four severity levels, it is difficult to get an overview of which ICD codes to focus on. To help users, all the unusable codes at each level were therefore grouped into ‘packages’ of CODs that had similarities and were given labels. For example, all the different misdiagnosis pathways that are coded to septicaemia are grouped into a ‘sepsis’ package; all diagnostic pathways that are incorrectly coded to urinary and genital fistula are grouped into the ‘fistula’ package, and so on. In all, there are some 160 such unusable code packages for ICD 10 coded causes across the four levels. Rather than dealing with a large number of individual unusable CODs, the packages can be ranked in order of importance within each of the four levels. As a result, it is possible to immediately see the leading packages of unusable CODs that cause the greatest problems within each severity level.

Within each of these packages, ANACONDA further offers the possibility of identifying the top 10 ICD 10 codes that are contributing the largest number of unusable codes to a specific package. It is this detailed information that is likely to be most useful in guiding COD improvement strategies. Countries should focus attention on the most important packages listed in levels 1–3 of the severity classification, and within these identify the key ICD codes that are being most misused. The packages and ICD 10 codes listed in level 4, although relevant for some countries, can be given lower emphasis in a national COD data improvement strategy.

In the second last substep (**7.8**), ANACONDA investigates whether there are any nonstandard ICD codes in the dataset. Some countries may, for very good reasons, create additional CODs. Because these cannot be classified to any ICD code, ANACONDA considers them as unknown causes and excludes them from the analysis. Finally, in **substep 7.9**, the application checks whether there are deaths that have been assigned a biologically implausible cause. Most of the biologically implausible causes arise from children being assigned adult causes such as lung or ovarian cancer, or males being assigned a maternal COD or coded as dying from strictly female cause such as ovarian cancer (and the converse). Neither of these two substeps is likely to constitute a significant proportion of the input data, but at least the implausible COD can be corrected once revealed.

## Step 8: Age pattern of mortality for broad disease groups

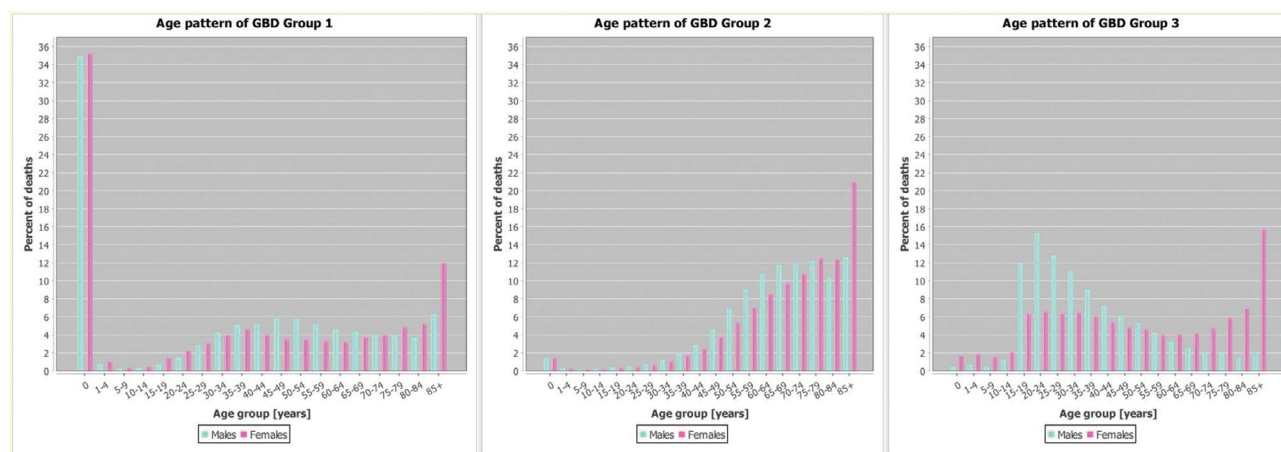
# 8



Most common CODs in a population follow a predictable age pattern that has been identified from decades of epidemiological observation. ANACONDA **substep 8.1** checks whether the age patterns for the three major disease and injury groups are consistent with what one would expect from this research knowledge.

Figure 16 shows age patterns for three major disease and injury groups that differ substantially. This is because the risk of dying from the different diseases and injuries covered in each group varies with age. For example, most deaths for from diarrhoea and malaria are children, while the victims of most circulatory diseases and cancers are older people. The age pattern of death from any given disease or injury should remain relatively the same, irrespective of life expectancy in a population, with a few well-known exceptions such as tuberculosis. What changes with increasing life expectancy is the proportion of all deaths in each major group (ie the epidemiological transition).

Figure 16: Age pattern of broad disease groups







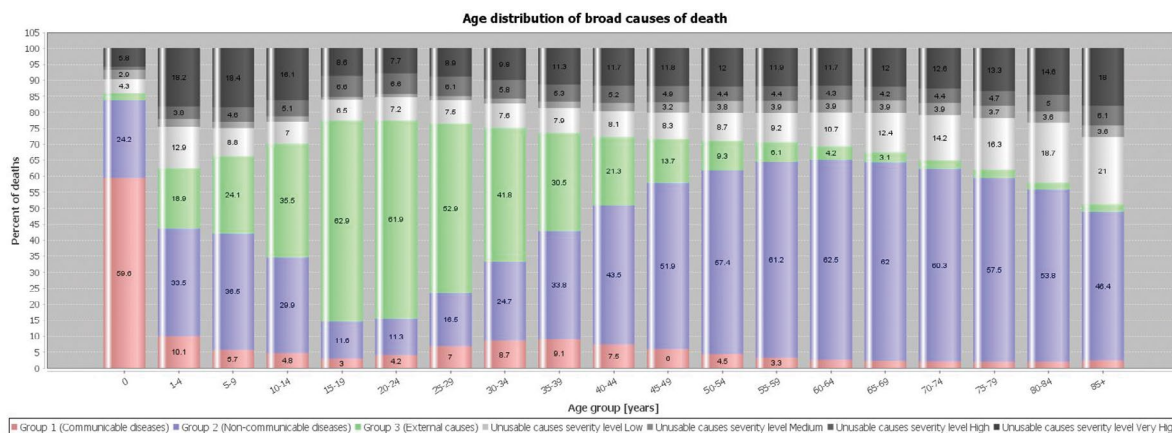
## Understanding substep 8.1

The typical pattern for group I (communicable diseases) is that they cause significant mortality in younger children, but thereafter decline to a very low level, only to rise again at older ages, usually above 80, as a result of low resistance to infectious diseases. Although group II causes (noncommunicable diseases) also contribute to some mortality in children, particularly due to congenital malformations, the vast majority of deaths from these diseases can be expected to occur at older ages, typically 50+. For group III (external causes), the proportion of deaths is generally highest in young adulthood, particularly for male deaths due to traffic accidents and violence. However, it can also be high at older ages, particularly for females, due to falls and other accidents.

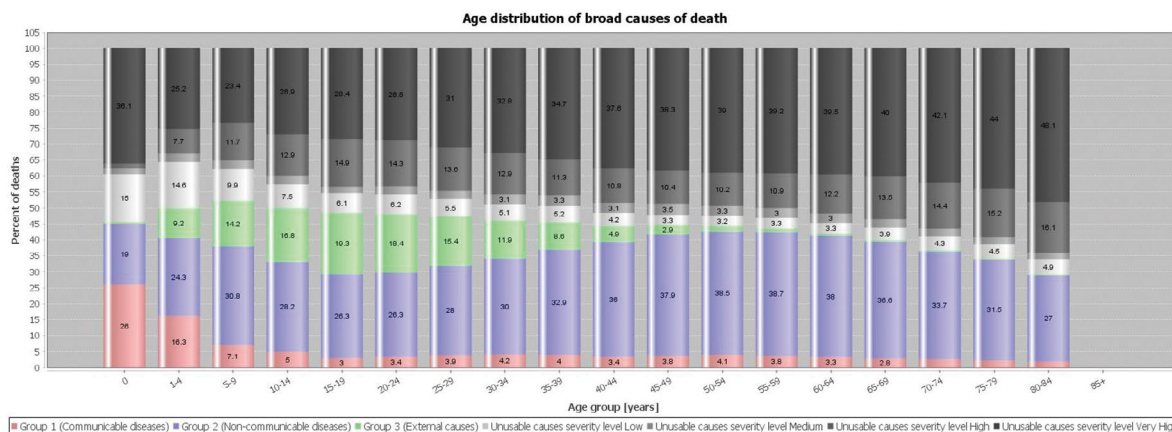
The analysis of **substep 8.2** brings the three patterns in Figure 16 together in one chart and adds further perspective by adding to that the age-specific unusable pattern. The two examples in Figure 17 show for each age group the relative importance of the three major disease groups, along with the proportion of cases where the true cause is unknown. Figure 17 also illustrates the degree of confidence or mistrust we can have in the COD data overall, but particularly at the oldest ages.

Figure 17: Examples of ANACONDA output - distribution of deaths by broad groups, unusable and insufficiently specified CODs in males and females by age group

### Example 1



### Example 2





## Understanding substep 8.2

The first chart in Figure 17 (example 1) shows, as expected, that the highest share of communicable diseases is found in children. For ages 10–30, the highest proportion die from external causes such as accidents, violence and suicides. From 40 years and onwards, most deaths are due to noncommunicable diseases. Only for the age groups above 75 years are there 10% or more deaths that are poorly specified. This indicates that this dataset can be trusted and used for guiding policy.

In the second chart (example 2), where we do not know for sure the COD of close to half of the deaths in each age group, it is still possible to distinguish the same patterns as described above. However, the extensive number of unusable and poorly specified codes leads to great uncertainty in what the true COD distribution might be, as it is unlikely that the unusable codes have a similar COD distribution across the three groups. This means that, even at this broad cause level, the data are not sufficiently reliable to be used for policy.

# 9



## Step 9: Leading causes of death

All health information systems should, as a bare minimum, be able to produce a table showing the leading CODs for the population. Such rankings are an important source of policy-relevant information. This information informs decisions on disease- and injury-control priorities and interventions to prevent premature mortality, and is useful for monitoring the impact of interventions. An analysis of leading CODs can also indicate the reliability and plausibility of the COD data.

**Step 9** investigates which causes, according to the input data, are the leading CODs in the country and whether these are plausible. It also ranks the 20 leading CODs in the input dataset for males and females. If any of these are considered unusable causes and belong to levels 1–3 in the severity classification in **substep 7.4** (ie they have very high, high or medium impact), they are highlighted in red to indicate that they are of no value to guide public policy and are a sign of poor certification practices. The fact that they appear among the 20 leading causes in the input dataset should be of concern. A light orange highlight on a leading cause indicates that there is a more specific ICD code for the disease and hence the COD can be considered as ‘insufficiently specified’ (the fourth level). While it would be desirable for certifying physicians to be more specific in assigning a disease subtype in such cases (where it is possible to do so), the consequences for disease reduction strategies are minor compared with the other three levels of unusable codes. For example, unspecified pneumonia diagnoses may hinder pneumonia control efforts if not further disaggregated, but the additional detail may not be essential for policy debates to improve population health (see Table 5).

**Table 5: Example ANACONDA output: leading COD, (with unusable (red) and insufficiently specified (orange) codes for COD highlighted)**

Rank	% of causes	ICD code	Name of category
1	11.9	I50.-	Heart failure
2	5.9	I46.-	Cardiac arrest
3	5.7	K72.-	Hepatic failure, not elsewhere classified
4	5.7	K74.-	Fibrosis and cirrhosis of liver
5	5.5	I61.-	Intracerebral haemorrhage
6	5.0	I10.-	Essential (primary) hypertension
7	2.7	R54.-	Senility
8	2.4	J96.-	Respiratory failure, not elsewhere classified
9	2.1	I74.-	Arterial embolism and thrombosis
10	2	I70.-	Atherosclerosis
11	1.8	I21.-	Acute myocardial infarction
12	1.7	R09.-	Other symptoms and signs involving the circulatory and respiratory systems
13	1.6	R73.-	Elevated blood glucose level
14	1.5	N17.-	Acute renal failure
15	1.5	J18.-	Pneumonia, organism unspecified
16	1.4	C22.-	Malignant neoplasm of liver and intrahepatic bile ducts
17	1.3	I63.-	Cerebral infarction
18	1.3	B19.-	Unspecified viral hepatitis
19	1.2	I64.-	Stroke, not specified as haemorrhage or infarction
20	1.2	N18.-	Chronic renal failure



### Understanding substep 9.1

The more frequently these unusable and insufficiently specified categories appear in the list of leading causes, and the higher they are ranked, the poorer is the quality of the input data and the more limited the data will be in providing essential health intelligence about the true picture of leading CODs in the population.

#### (4) Overall data quality index: VSPI(Q)

# 10



### Step 10: Vital Statistics Performance Index

ANACONDA offers a large number of indicators to flag various quality aspects of mortality and COD data. One single summary score of the overall performance of a CRVS system is useful for monitoring CRVS system improvement. Such a score should take into account all the essential components of data quality, and weight them according to their importance in affecting the overall utility of the data for policy development.

In this step, ANACONDA automatically calculates a summary index of mortality data quality, known as the VSPI(Q). The VSPI is based on a broader metric of CRVS data utility developed by Philips et al (2014), which also includes a component to reflect the timeliness of the data. This component is not included in the VSPI(Q), which focuses on the quality of the input data, and assesses CRVS system performance against five key measures of data quality, all derived from the input data:

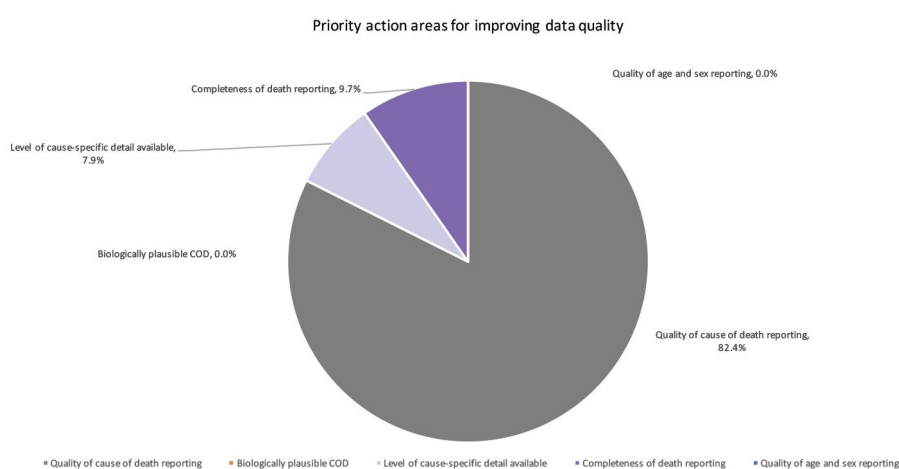
1. **completeness of death registration** (as predicted from the input data; see **substep 2.2**)
2. **fraction of unusable codes in the data** (based on the four-level classification of unusable codes according to their implications for guiding policy, with levels 1, 2 and 3 given twice as much weight as level 4 in the overall score for this component)
3. **amount of detail in the COD list** used to tabulate data
4. **extent to which age and/or sex are missing** in the data
5. **number of biologically implausible underlying causes.**

Scores on each of these five components are then weighted according to their importance in determining the correct COD distribution in a population, and combined into a VSPI(Q) score, ranging from 0 to 100. The higher the score, the better the overall quality of the mortality data.

The VSPI(Q) calculated in this way does not require any additional input information to what is already required as input into ANACONDA, at least for national applications. The VSPI(Q) can also be readily calculated at the subnational. In in this case, given the likelihood that mortality levels will vary across a country, it is recommended that differences in the level of child mortality in different regions be included as input into the model for predicting completeness of death registration (see **step 2**).

Figure 19: Example ANACONDA output – the Vital Statistics Performance Index for Data Quality

<b>Country:</b>	
<b>Region:</b>	7 (Latin America/ Caribbean)
<b>Summary score:</b>	78.10%
<b>Classification:</b>	High
<b>VSPI Quality Components Score</b>	
<b>Component</b>	<b>Score (weighted)</b>
Quality of cause of death reporting	81.3
Completeness of death reporting	97.8
Level of cause-specific detail available	98.2
Quality of age and sex reporting	100.0
Biologically plausible COD	100.0





## Understanding step 10

Figure 19 shows the output from step 10 in ANACONDA. The panel on the left gives the overall VSPI(Q) score, as well as an evaluation of the score in terms of five performance categories (*very low, low, medium, high, very high*) (Mikkelsen et al 2015). The weighted scores for each of the five components of the index are also shown. These individual component scores are useful for pinpointing the most important failings that affect the system and hence for establishing priorities for action. Since the component scores are simply multiplied to yield the overall VSPI(Q) score, they are often individually higher than the overall score.

The pie chart in Figure 19 shows the specific contribution of each component (expressed as a percentage) to the gap between the VSPI(Q) score and 100%. The larger the share a component has of the pie, the more important that component is in bringing down the overall VSPI(Q) score for the country, and the more urgent are efforts to improve it to increase the overall quality of the input data.



## Summary

ANACONDA is a simple, yet very informative, tool for countries to use to quickly check the quality of the data they routinely collect on mortality and CODs. Timely and accurate COD statistics have enormous policy value, so it is important that governments understand the likely biases and errors in their routine mortality data collections, and take urgent and systematic steps to correct them. ANACONDA offers a structured, sequential framework to help countries identify the main problems and errors in their mortality data. At the same time, it serves as a teaching tool to train data producers and analysts to think critically about data quality, and provides them with the evidence needed to carry out improvement actions.

This guidance document details the logic and rationale underlying a systematic approach to interrogating mortality statistics to assess data quality issues. It gives users of ANACONDA an in-depth explanation of the different components of the framework, why they were included, what they tell us about the data and how to interpret the results.

Large datasets generated for public health purposes will never be error-free. There will always be deaths that will be classed as ill-defined, because the underlying cause cannot be identified. This is more likely to be the case for the older ages, when multiple morbid conditions are often present around the time of death, and correctly deciding on the sequence of events that led to death is difficult, if not impossible. Death certification is not a precise science, but one where a physician is asked to apply their clinical knowledge to determine the underlying cause that led to death. In many cases, the physician may not have been the treating physician and has little or no knowledge of the clinical history of the deceased. Yet, even imperfect COD data can provide some information for policy if analysts can apply the methods outlined in the ANACONDA tool to assess the probable extent and patterns of biases in the data.


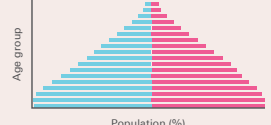

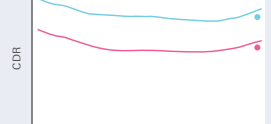

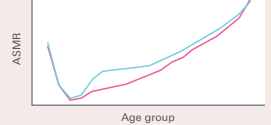

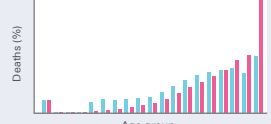

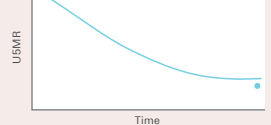





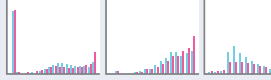




By regular applications of ANACONDA, and careful interpretation of the outputs as suggested in this guide, we expect that users will soon develop a culture of critical appraisal of mortality statistics. They will gain the confidence and competence needed to interpret their routine mortality datasets according to this guide's comprehensive processes and principles. Apart from evaluating how reliable the input mortality dataset is, ANACONDA will very clearly highlight the areas where the data are flawed and in need of urgent intervention to rapidly strengthen health information and civil registration systems.

Some of the interventions that are likely to be required to strengthen existing systems and generate better mortality data are described by Mikkelsen et al (2016). Foremost among these is the need to apply a systematic and strategic approach to training doctors in the correct certification of CODs using the latest International Form of the Medical Certificate of Causes of Death recommended by WHO. Doctors often do not receive adequate training in the correct completion of the COD certificate, and often do not appreciate the critical public health importance of the information compiled from these certificates. Training strategies for medical students or recent medical graduates might have to be carefully evaluated, and strengthened as necessary. The primary aim of such strategies is to dramatically reduce the extent of unusable or unspecified COD codes in the data, and thus substantially improve the value of the data for national public health programs and policies.

The other set of interventions that ANACONDA will likely indicate are those designed to strengthen death registration practices, including adopting more efficient and equitable procedures for registering deaths, and for ensuring that the critical information about each death, including the precise age and sex of the decedent, is correctly recorded. Subnational application of the tool will help to identify those sectors of the population where registration completeness is particularly poor. In all cases, it is likely that death registration will be lowest among children, indicating the need for a careful review of current registration practices and disincentives.

Finally, it is important to appreciate that rapid change in the quality of COD statistics is possible, if there is firm government commitment to improve data quality and to use the data for informing public policy. Several countries, including Turkey, South Africa and Brazil, have demonstrated that significant change is possible within a short time with targeted application of specific CRVS strengthening methods (Mikkelsen et al 2015). ANACONDA will help countries identify the need for these strategies, and provides a convenient means of monitoring progress with them, particularly if it is integrated into the annual routine data production process.

## ANNEX 1: ANACONDA's 10 steps

<p>1 </p>		<p><b>Data input checks</b></p> <p>Provides an overview of the input data that allows you to check for any potential errors or inconsistencies. Tabulates deaths by standard International Classification of Diseases (ICD) and Global Burden of Disease (GBD) tabulation lists, by age and sex.</p>
<p>2 </p>		<p><b>Crude death rate</b></p> <p>The estimated and calculated crude death rates (CDR) from the input data are used to assess the extent of potential underreporting of deaths.</p>
<p>3 </p>		<p><b>Age-specific mortality rates</b></p> <p>The age- and sex-specific mortality rates are shown in a log plot. Inconsistencies such as a nonlinear line after age 35 should be investigated as they could indicate incompleteness of death reporting. The male rates should be consistently higher for all ages, especially between 20 and 35 years of age.</p>
<p>4 </p>		<p><b>Age distribution of deaths</b></p> <p>The age distribution of deaths should show a higher concentration of deaths among children under 1 year of age, be lowest at ages 5–14, be followed by a rapid increase for males in their early twenties, and then gradually increase with age for males and females.</p>
<p>5 </p>		<p><b>Completeness of child mortality</b></p> <p>This step compares the calculated level of child mortality from the input data with external estimates from censuses and surveys, allowing you to calculate the relative difference between the two. This gives an estimate of the extent of underregistration of child deaths. This step also produces a life table from the input data, which includes life expectancy.</p>
<p>6 </p>		<p><b>Mortality by broad GBD groups</b></p> <p>An important first step in assessing the quality of cause of death (COD) data is to look at the distribution of deaths by three broad cause groups (communicable, noncommunicable, external) and assess whether they are consistent with expected patterns given current mortality conditions. This step also shows the number of deaths assigned to unusable and insufficiently specified ('garbage') causes, which is an important indicator of data quality.</p>
<p>7 </p>		<p><b>Quality of cause of death data</b></p> <p>This step analyses the extent of COD diagnoses in the input data that are of no or limited use because they do not accurately reflect the true underlying COD. The unusable causes of death are further classified into types of errors, and into severity levels according to the impact they can have on misguiding policy and planning.</p>
<p>8 </p>		<p><b>Age pattern of mortality by broad groups</b></p> <p>As the risk of dying from different diseases and injuries changes with age, the age pattern of deaths within each of the three broad cause groups will also be different. If you do not see a distinct age pattern for each of these three groups, you are likely to have problems with misdiagnosis in the input data.</p>
<p>9 </p>		<p><b>Leading causes of death</b></p> <p>A useful way to gain an overview of the policy utility of mortality data is to rank the leading COD. There should be no unusable causes (highlighted in red or orange) ranked among the 20 leading causes of death.</p>
<p>10 </p>		<p><b>Vital Statistics Performance Index</b></p> <p>The Vital Statistics Performance Index (VSPI) is a single summary score of the performance of a vital statistics system, which takes into account five essential components of data quality. The overall VSPI score can be broken down into scores for each of the five different components. The lower the component scores (bigger segments of the pie chart), the higher priority should be given to that component in strategies to improve the data.</p>

## ANNEX 2: Model for estimating completeness

Step 2 is designed to check the possible extent of underregistration of deaths across all ages and both sexes. This is done in two ways:

1. First, by calculating the crude death rate (CDR, defined as total deaths total midyear population x 1000) from the input data and comparing it with estimates of the actual level of the CDR for the same country-year from the Global Burden of Disease (GBD) study. A CDR <5 in any population where the risk of child mortality (ie risk of death before age 5; see step 5) is >30/1000 live births should be viewed suspiciously and is likely to be indicative of underregistration of deaths in that dataset.
2. Second, by using a new method that models the level of the reported CDR for a given population as a function of the key drivers of the CDR, namely population age composition and the overall level of mortality in a population (as approximated by the level of child mortality).<sup>8</sup> The model (in logit space) is as follows:

### Model equation:

$$\text{logit}(Completeness) = (\text{RepCDR2} * -0.0187471) + (\text{RepCDR} * 0.6125569) + (\%65plus * -12.58245) + (\ln(5q0) * -1.134923) + (\text{Compl}_5q0 * 2.319505) + (\text{Year} * -0.0184299) + 31.40303 + \gamma$$

where:

- *Completeness* is the estimated proportion of all deaths in the population that are registered or otherwise notified (ie. it is a fraction)

$$Completeness = \frac{e^{\text{logit completeness}}}{1 + e^{\text{logit completeness}}} \text{ and } \text{logit}(Completeness) = \ln\left(\frac{Completeness}{1 - Completeness}\right)$$

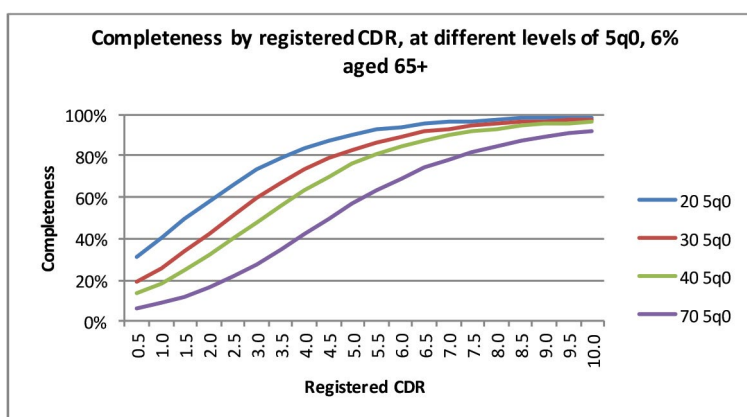
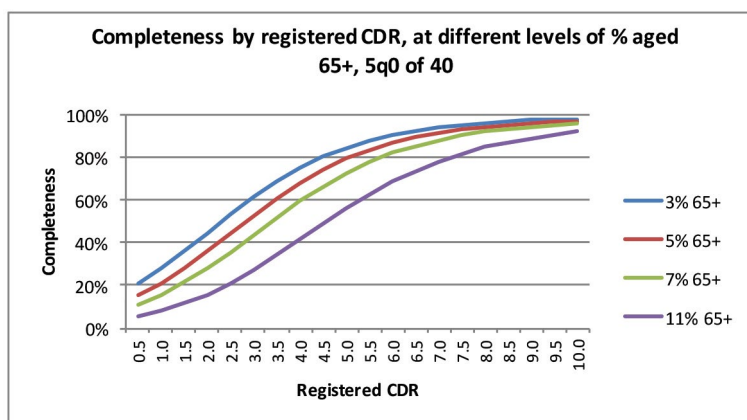
- *RepCDR* is the CDR calculated from the input data
- *%65plus* is the fraction of the input population aged 65 and over
- *ln(5q0)* is the natural log of the probability of child death before age 5 (per 1000 live births), used as an overall measure of mortality levels in the population
- *Compl\_5q0* is the estimated proportion of under-5 deaths in the population that are registered or notified
- Year is the calendar year of the data
- $\gamma$  is a country-level random effect. Each country that was used in the data to develop this statistical model has a random effect, which is a number that adjusts the completeness estimate based on systematic differences in completeness between countries that are not captured by the input data.

It is important that the level of  ${}_5q_0$  used in the model is as close as possible to the true level of child mortality in the population; this will generally mean that the rate calculated from the life table in step 5.2 cannot be used for this purpose. Rather, we use the level from the comparator (Inter-agency Group for Child Mortality Estimation) or a value input by the user based on censuses and surveys. To calculate the completeness of under-5 deaths, ANACONDA automatically uses the data in step 5.1 (ie the  ${}_5q_0$  from the registration data divided by the  ${}_5q_0$  from the comparator).

The graphs below show the modelled relationships between CDR (calculated from input data) and predicted completeness of death registration or notification in the population at different levels of child mortality and population aged 65+. These levels (ie  ${}_5q_0$  of 40/1000 and 6% aged population) have been chosen as examples to illustrate how completeness is likely to vary according to commonly observed values of  ${}_5q_0$  and percentage 65+.

<sup>8</sup> Adair T & Lopez AD (2017). *Estimating the completeness of death registration: an empirical method*. CRVS technical outcome series. University of Melbourne, Civil Registration and Vital Statistics Improvement Group, Melbourne. [Draft document in development].





This modelled relation is more applicable to estimating levels of death registration/notification completeness less than 90%; above that level of completeness, the CDR can (and does) vary significantly, often from 5/1000 to 12/1000, as a result of differences in the extent of population ageing.



## References

- AbouZahr C, Mikkelsen L, Rampatige R & Lopez AD (2010). *Mortality statistics: a tool to improve understanding and quality*. Working Paper no. 13, University of Queensland School of Population Health, Health Information Systems Knowledge Hub, Brisbane.
- Institute for Health Metrics and Evaluation (2015). Global burden of disease cause list 2015. [www.healthdata.org/sites/default/files/files/Projects/GBD/GBDcause\\_list.pdf](http://www.healthdata.org/sites/default/files/files/Projects/GBD/GBDcause_list.pdf)
- Lozano R et al (2013). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the Global Burden of Disease Study 2010. *Lancet*, volume 380(9859), pp. 2095–128.
- Mikkelsen L et al (2015). A global assessment of civil registration and vital statistics systems: Monitoring data quality and progress. *Lancet*, volume 386(10001), pp. 1395–1406.
- Mikkelsen L, Rampatige R, Hernandez B & Lopez AD (2016). *Strengthening vital statistics systems*, Policy Brief volume 1(2), Asia Pacific Observatory on Health Systems. [www.wpro.who.int/asia\\_pacific\\_observatory/resources/policy\\_briefs/strengthening\\_vital\\_statistics\\_systems/en](http://www.wpro.who.int/asia_pacific_observatory/resources/policy_briefs/strengthening_vital_statistics_systems/en)
- Murray CJL and Lopez AD (eds). 1996. *The Global Burden of Disease: a comprehensive assessment of mortality and disability from diseases, injuries and risk factors in 1990 and projected to 2020*. Global Burden of Disease and Injury Series, Vol. I. Cambridge, MA: Harvard University Press on behalf of the World Health Organization and The World Bank.
- Murray CJL et al (2007). Estimating population cause-specific mortality fractions from in-hospital mortality: Validation of a new method. *PLoS Medicine*, volume 4(11), p. e326.
- Murray CJL et al (2010). What can we conclude from death registration? Improved methods for evaluating completeness. *PLoS Medicine*, volume 7(4), p. e1000262. <https://doi.org/10.1371/journal.pmed.1000262>
- Phillips DE, Lozano R, Naghavi M et al (2014). A composite metric for assessing data on mortality and causes of death: The vital statistics performance index. *Population Health Metrics*, volume 12(14). [www.ncbi.nlm.nih.gov/pmc/articles/PMC4060759/](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4060759/)
- Rampatige R et al (2014). Hospital cause-of-death statistics: what should we make of them? *Bulletin of the World Health Organization*, volume 14(92), pp. 3–3A. [www.ncbi.nlm.nih.gov/pmc/articles/PMC3865553/](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3865553/)
- United Nations (2015). *Demographic components of future population growth: 2015 revision*, United Nations, New York. [www.un.org/en/development/desa/population/theme/trends/dem-comp-change.shtml](http://www.un.org/en/development/desa/population/theme/trends/dem-comp-change.shtml)
- University of Melbourne & Swiss Tropical and Public Health Institute (2017). *ANACONDA (V2.2) user guide*. University of Melbourne, Civil Registration and Vital Statistics Improvement Group, Bloomberg Philanthropies Data for Health Initiative, Melbourne. [http://mspgh.unimelb.edu.au/\\_data/assets/pdf\\_file/0018/2330091/HR\\_RT10\\_ANACONDA-user-guide\\_031701.pdf](http://mspgh.unimelb.edu.au/_data/assets/pdf_file/0018/2330091/HR_RT10_ANACONDA-user-guide_031701.pdf)
- University of Melbourne (2017). *ANACONDA: Facilitator's guide*. University of Melbourne, Civil Registration and Vital Statistics Improvement Group, Bloomberg Philanthropies Data for Health Initiative, Melbourne.
- Wang H et al (2016). Global, regional, and national life expectancy, all-cause specific mortality for 249 causes of death, 1980–2015: A systematic analysis for the Global Burden of Disease Study 2015. *Lancet*, volume 388, pp. 1459–544.
- World Health Organization (2016). *International statistical classification of diseases and related health problems*, 10th revision, vol. 2, 10th edn, World Health Organization, Geneva.



The program partners on this initiative include: The University of Melbourne, Australia; CDC Foundation, USA; Vital Strategies, USA; Johns Hopkins Bloomberg School of Public Health, USA; World Health Organization, Switzerland.

Civil Registration and Vital Statistics partners:



## For more information contact:

CRVS-info@unimelb.edu.au  
crvsgateway.info

CRICOS Provider Code: 00116K

Version: 1017-01

### Copyright

© Copyright University of Melbourne October 2017.

The University of Melbourne owns the copyright in this publication, and no part of it may be reproduced without their permission.

### Disclaimer

The University of Melbourne has used its best endeavours to ensure that the material contained in this publication was correct at the time of printing. The University gives no warranty and accepts no responsibility for the accuracy or completeness of information and the University reserves the right to make changes without notice at any time in its absolute discretion.

### Intellectual property

For further information refer to: [unimelb.edu.au/governance/statutes](http://unimelb.edu.au/governance/statutes)