Supplementary table S1

Ten data quality steps of ANACONDA

Step		Description
1	Data input checks	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
2	Crude death rate	The estimated and calculated crude death rates (CDR) from the input data are used to assess the extent of potential underreporting of deaths.
3	Age-specific mortality rates	The age- and sex-specific mortality rates are shown in a log-plot. Inconsistencies such as a non-linear 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.
4	Age-sex distribution of deaths	The age distributions of deaths should show a higher concentration of deaths among children under one-year of age, lowest at ages 5-14, followed by a rapid increase for males in the their early twenties, and then gradually increasing with age for males and females.
5	Completeness of child mortality	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 under-registration of child deaths. This step also produces a life table from the input data, which includes life expectancy
6	Mortality by broad GBD groups	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; non-communicable; 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.
7	Quality of cause of death data	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.

The ICD-10 groups mortality codes into 22 broad chapters. By displaying the proportion of deaths belonging to each chapter of the ICD-10, and the fraction of unusable and insufficiently specified codes within each chapter, it is possible to immediately see where these codes are coming from and the major areas of concern. ANACONDA analyses the unusable and insufficiently specified codes and classifies these into five different categories according to ICD concepts, as follows:

- *Category 1:* Codes relating to symptoms, signs and ill-defined conditions (mostly drawn from R00–R99 in ICD-10).
- *Category 2:* Codes that have an impossible underlying cause of death.
- *Category 3:* Codes relating to intermediate causes of death.
- *Category 4:* Codes relating to immediate causes of death, such as heart or respiratory failure.
- *Category 5:* Insufficiently specified codes within ICD chapters within a larger disease category.

ANACONDA also provides an alternative approach to classifying unusable and insufficiently specified codes. This second classification regroups the unusable and insufficiently specified codes according to their potential impact for guiding or misguiding public policy to prevent premature deaths. These four levels are defined as:

- **Level 1 – Codes with serious implications likely to have a *very high impact*** for health policy. These are codes relating to such vague causes, that the true underlying cause of death could belong to more than one broad cause group.
- **Level 2 – Codes with substantial implications likely to have *a high impact***. These are codes relating to vague causes, where the true cause of death is likely to belong to only one of the three broad groups.
- **Level 3 – Codes with important implications likely to have *medium impact***. These are codes for which the true underlying cause of death is known to be within the same ICD chapter. For instance, a death assigned to 'ill-defined site of cancer' indicates that

		the true cause of death was cancer but does not specify the site. • **Level 4 – Codes with limited implications likely to have *low impact***. These are codes for which the true cause of death 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 this level will generally be minor. Given the thousands of ICD-10 codes and the large number that are no use for mortality analysis and public health ANACONDA groups the unusable mortality codes into the 'packages' of similar unusable codes within each of the four levels of severity. Each package groups together the codes resulting from poor diagnostic practices and coding, resulting in specific, identifiable misdiagnoses. The utility of this is that for each case the specific code can be identified and be the focus of correction efforts.
8	Age pattern of mortality by broad groups	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.
9	Leading causes of death	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.
10	Vital Statistics Performance Index (VSPI)	The Vital Statistics Performance Index for Quality, or VSPI(Q), is a summary score of overall system performance that takes into account five essential components of quality:[1] 1. Completeness of death registration 2. Quality of cause of death reporting (fraction of unusable or insufficiently specified codes) 3. Level of cause-specific detail available (amount of detail in the cause of death list used for tabulation) 4. Quality of age and sex reporting (extent to which age and/or sex are missing in the data)

[1] Philips DE, Lozano R, Naghavi M, et al. A composite metric for assessing data on mortality and causes of death: the vital statistics performance index. *Population Health Metrics* 2014: 12(14).

		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 cause of death 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, with values above 85% suggesting a well-functioning CRVS system that will meet most policy needs for reliable data.