Sampling Strategies for Representative National CRVS Verbal Autopsy Planning:

A Guidance Document and Sample Size Calculator Tool

Part A: Principles and Strategy

Version 2.4

July 26, 2018

Review Version

Contents

Part A: Principles and Strategy **1**

Acronyms **6**

Preface **7**

Acknowledgements **9**

Executive Summary **10**

Part A - Principles and Strategy **13**

1. Introduction **13**

1.1 Pathways to scale for CRVS Verbal Autopsy 13

1.2 Rationale for CRVS VA sampling 14

1.3 Rationale for cluster sampling 14

2. Key principles for National CRVS VA Sampling **15**

2.1 Non-competition with medical certification 15

2.2 The need for effective universal death notification and registration 16

2.3 Deaths without medical certification of cause of death 16

3. Strategic operational considerations for National CRVS VA Sampling **17**

3.1 Defining the operational cluster 17

3.2 How many Cause Specific Mortality Fractions? 17

3.3 Disaggregation of results 17

3.3.1 Male-Female disaggregation 18

3.3.2 Age Group disaggregation 18

3.3.3 Urban-rural disaggregation 18

3.3.4 Sub-national administrative disaggregation 19

3.4 De-duplication 19

4. Considerations for framing the CRVS VA sample **19**

4.1 The Sample Frame 19

4.2 Inclusions and exclusions 20

5. Considerations for calculating the CRVS VA sample size **20**

5.1 Statistical approach used in the CRVS VA Sample Size Calculator Tool 20

5.2 Overview of the CRVS VA Sample Size Calculator Tool 20

5.3 Tool input parameters required 21

5.3.1 Number of clusters 21

5.3.2 Maximum acceptable uncertainty range 21
5.3.3 Mean cluster population 21
5.3.4 Crude death rate 22
5.3.5 Number of years to aggregate for trend 22
5.3.6 Percentage of deaths with MCCD 22
5.3.7 Under-notification and non-response rate 22
5.3.8 Scenario where population size of eligible clusters is known 23
5.4 Tool output parameters produced 24
5.4.1 Number of clusters required 24
5.4.2 Estimated total population in the sample 24
5.4.3 Estimated number of deaths in the sample per year 24
5.4.4 Estimated number of VAs needed per year 24
5.4.5 CSMF uncertainty ranges 25
6. Considerations for selecting the CRVS VA sample clusters 25
6.1 Defining the sample selection strategy 25
6.1.1 Stratification 25
6.1.2 Simple random sampling 26
6.1.3 Systematic sampling 26
6.1.4 Probability Proportional to Size sampling 26
6.1.5 Stratified single-stage cluster PPS sampling 26
6.2 Documenting the sample method 27
7. Limitations 27
8. Scaling up strategy and follow-up period 28
9. Conclusions 28
Part B: Methods and Tools 30
Glossary 31
1. Introduction 34
2. Preparing the sampling frame 34
3. Calculating the sample size 35
3.1 Preparatory steps 35
3.2 Using the CRVS VA Sample Size Tool 36
4. Selecting the sample clusters 38
4.1 Stratification 38
4.2 Probability Proportional to Size sampling 38

ANNEXES 40

Annex A. Statistical basis of the CRVS VA sample size calculations 41

A.1. Assumptions 41

A.2. Power and Significance level 41

A.3. Individual vs. Cluster design 41

A.4. Unmatched vs. Matched design 41

A.5. Design related parameters 42

A.5.1. Coefficient of variation between clusters 42

A.5.2. Intra-cluster correlation coefficient 42

A.5.3. Design effect 42

A.5.4. Coefficient of variation in cluster size 43

A.6. Formula for sample size based on proportions 43

A.7. Formula for sample size based on rates 43

A.8. Cluster size 43

A.9. Uncertainty range 44

A.10. Further Adjustments 44

A.10.1. Disaggregation by male and female 44

A.10.2. Proportion of deaths having MCCD 44

A.10.3. Under-notification and non-response rate 44

Annex B. Worked example implementing the Guidance and Tool in one country 45

B.1. Preparing to calculate the Cluster Sample Size 45

B.1.1. Operational cluster definition 45

B.1.2. Disaggregation level of results 45

B.1.3. Number of years to aggregate for trend 45

B.1.4. National level estimates 45

B.2. The sampling frame 48

B.3. Estimating k 49

B.3.1. Options to estimate k 49

B.4. Estimating MIS 53

B.5. Calculating the CRVS VA sample size 54

B.6. Sampling strategy for selecting the CRVS VA sample clusters 58

B.7. Calculating the number of clusters required in order to disaggregate results for male and female 60

B.8. Calculating the number of clusters required separately for areas with different proportions of deaths with an MCCD 62

Annex C. Additional Resource Material 64

C.1. National Crude Death Rate estimates for Data for Health Initiative CRVS VA countries 64

C.2. Method for estimating national level CSMFs 65

C.3. Top 20 CSMF estimates for Data for Health Initiative CRVS VA countries 67

C.4. Verbal Autopsy Target Cause Lists 76

C.5. Link to CRVS VA Costing Tool for download 76

References 78

Acronyms

CDR	Crude Death Rate
CRVS	Civil Registration and Vital Statistics
CSMF	Cause-Specific Mortality Fraction
CSMR	Cause-Specific Mortality Rate
CV	Coefficient of variation in cluster size
DE	Design Effect
DHIS2	District Health Information System 2
GIS	Geographic Information System
HDSS	Health and Demographic Surveillance System
ICC	Intra-cluster Correlation Coefficient
INDEPTH	International Network for Demographic Evaluation of Populations and their Health
k	Coefficient of variation of the true outcome measure between clusters at one point in time
k_m	Coefficient of variation of the true outcome measure between clusters within the matched pairs in absence of anything which could change the mortality and/or the CSMFs
MCCD	Medical Certification of Cause of Death (sometimes MCCOD)
MIS	Maximum possible Inflation in sample Size
PPS	Probability Proportional to Size sampling
RS	Random Start
SAVVY	Sample vital events with verbal autopsy
SCI	Symptom-Cause Information
SI	Sampling Interval
SMoL	Start-up Mortality List for ICD Coding
SRS	Sample Registration System
VA	Verbal Autopsy
WHO	World Health Organization

See Part 2, Methods and Tools, for a Glossary defining these and other terms used throughout this Guidance Package.

Preface

The purpose of this package of a two part Guidance Document and its companion Sample Size Calculator Tool is to assist countries with scale up and rollout planning for the application of verbal autopsy (VA) as a function of a national Civil Registration and Vital Statistics (CRVS) system.

Users of the package will include those tasked with designing and managing the CRVS VA system. The package is intended to be used after pre-testing and pilot phases during which the processes, methods and possibly costing of the CRVS VA system are perfected and established, and before the scale-up and rollout phase. Every country will have different implementation circumstances. Therefore, this document is necessarily generic in hopes that the considerations, options, and methods provided can be adapted and adjusted to the majority of circumstances.

As of 2018, 13 countries participating in the Bloomberg Philanthropies Data for Health Initiative are engaged in a pre-test or pilot phase implementation of mobile automated VA. Automated VA is intended to be integrated into their CRVS systems to improve availability of cause of death data for deaths without a medically certified cause, most of which occur in the community. These initial pre-tests or pilots, usually at district or equivalent scale, have allowed countries to start to cost out and plan national scale implementation. Although there has been limited experience with sample vital event "registration" systems with VA (China, India, Indonesia, Tanzania, and Zambia) these have not been fully integrated with CRVS and do not act to officially register vital events.

In cases where national CRVS authorities wish to take a representative sampling approach to VA implementation, there remain several open questions and a lack of practical guidance for how to estimate the annual number of VAs needed to provide valid and representative cause-specific mortality fractions (CSMFs). These data play an important role in forming health policy and program decisions. There are also questions about the issue of national and sub-national stratification. Stratification may be needed to address possible disparities due to ethnic, socio-economic, demographic (e.g. urban/rural) and epidemiologic factors. Therefore, a diverse expert group was convened to deliberate on these issues and prepare a practical guidance document and tool to assist countries scaling up from pre-test or pilot phases to national CRVS VA systems. This Technical Guidance Document and associated tool is the result.

Development of Concepts:

Initial discussions on the concept of CRVS VA sampling strategies were held as part of an International Consultation Workshop organized and financed by the Bloomberg Philanthropies Data for Health Initiative and convened at the Swiss Tropical and Public Health Institute CRVS Innovation Hub in Basel Switzerland, August 18-19, 2017. Participants at this workshop included:

Alan Lopez, Bloomberg Philanthropies Data for Health Initiative, University of Melbourne, Australia.
Daniel Chandramohan, Department of Disease Control, London School of Hygiene & Tropical Medicine, London, UK.
Daniel Cobos, Bloomberg Philanthropies Data for Health Initiative, CRVS Innovation Hub, Swiss Tropical and Public Health Institute, University of Basel, Basel, Switzerland.
Deidre McLaughlin, Bloomberg Philanthropies Data for Health Initiative, University of Melbourne, Melbourne, Australia.
Don de Savigny, Bloomberg Philanthropies Data for Health Initiative, CRVS Innovation Hub, Swiss Tropical and Public Health Institute, University of Basel, Basel, Switzerland.
Erin Nichols, Bloomberg Philanthropies Data for Health Initiative, NCHS US Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, Maryland, USA.

Gregory Kabadi, Bloomberg Philanthropies Data for Health Initiative CRVS Country Coordinator, Dar es Salaam, Tanzania.
Jordana Leitao, WHO Geneva, Verbal Autopsy Working Group, Geneva, Switzerland.
Magdalena Paczkowski, Bloomberg Philanthropies Data for Health Initiative, Vital Strategies, New York, NY, USA.
Maigen Zhou, China CDC, Shanghai, China.
Margarita Ronderos, Bloomberg Philanthropies Data for Health Initiative CRVS Technical Advisor, Bogota, Colombia.
Martin Bratschi, Bloomberg Philanthropies Data for Health Initiative, Vital Strategies, Singapore.
Peng Yin, China CDC, Shanghai, China.
Prasanta Mahapatra, Institute of Health Systems, Hyderabad, India.
Sam Clark, Ohio State University, Westerville, Ohio, USA.
Soewarta Kosen, National Institute of Health Research and Development, Jakarta, Indonesia.
Tom Smith, Swiss Tropical and Public Health Institute, Infectious Disease Modeling Unit, University of Basel, Basel, Switzerland.

Preparation of the CRVS VA Sampling Strategy Guidance Document:

The first drafts of Part A of this guidance document were prepared by Don de Savigny. Sabine Renggli prepared the first drafts of Part B and all annexes. All others below have contributed significantly to subsequent drafts of the combined document and final product.

Adam Karpati, Bloomberg Philanthropies Data for Health Initiative, Vital Strategies, New York. USA.
Daniel Cobos, Bloomberg Philanthropies Data for Health Initiative, CRVS Innovation Hub, Swiss Tropical and Public Health Institute, University of Basel, Basel, Switzerland.
Don de Savigny, Bloomberg Philanthropies Data for Health Initiative, CRVS Innovation Hub, Swiss Tropical and Public Health Institute, University of Basel, Basel, Switzerland.
Erin Nichols, Bloomberg Philanthropies Data for Health Initiative, US Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, Maryland, USA.
Martin Bratschi, Bloomberg Philanthropies Data for Health Initiative, Vital Strategies, Singapore.
Philip Setel, Bloomberg Philanthropies Data for Health Initiative, Vital Strategies, New York. USA.
Sabine Renggli, Bloomberg Philanthropies Data for Health Initiative, CRVS Innovation Hub, Swiss Tropical and Public Health Institute, University of Basel, Basel, Switzerland.
Sam Notzon, Bloomberg Philanthropies Data for Health Initiative, US Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, Maryland, USA.

Development and testing of the CRVS VA Sample Size Calculator Tool:

This guidance package includes an associated user-friendly CRVS VA Sample Size Calculator Tool in MS Excel. The following individuals contributed to the conceptualization, design, development and testing of the tool or contributed to the documentation of the tool.

Christian Schindler, Biostatistics Unit, Swiss Tropical and Public Health Institute, University of Basel, Switzerland.
Daniel Chandramohan, Department of Disease Control, London School of Hygiene & Tropical Medicine, London, UK.
Daniel Cobos, Bloomberg Philanthropies Data for Health Initiative, CRVS Innovation Hub, Swiss Tropical and Public Health Institute, University of Basel, Basel, Switzerland.
Don de Savigny, Bloomberg Philanthropies Data for Health, CRVS Innovation Hub, Swiss Tropical and Public Health Institute, University of Basel, Basel, Switzerland.

Erin Nichols, Bloomberg Philanthropies Data for Health Initiative, US Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, Maryland, USA.
Gregory Kabadi, Bloomberg Philanthropies Data for Health Initiative CRVS Country Coordinator, Dar es Salaam, Tanzania.
Hee-Choon Shin, US Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, Maryland, USA.
Isaac Lyatuu, Department of Epidemiology and Public Health, Swiss Tropical and Public Health Institute. University of Basel, Switzerland.
Jon Wakefield, University of Washington, Seattle, WA, USA.
Katherine Fielding, Medical Statistics and Epidemiology, London School of Hygiene & Tropical Medicine, London, UK.
Lea Multerer, Department of Epidemiology and Public Health, SwissTPH. University of Basel, Switzerland
Philip Setel, Bloomberg Philanthropies Data for Health Initiative, Vital Strategies, Seattle, WA. USA.
Richard Hayes, London School of Hygiene & Tropical Medicine, London, UK.
Sabine Renggli, Bloomberg Philanthropies Data for Health Initiative, CRVS Innovation Hub, Swiss Tropical and Public Health Institute, University of Basel, Switzerland.
Sam Clark, Ohio State University, Westerville, Ohio, USA.
Yulei He, US Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, Maryland, USA.

Acknowledgements

This work was supported financially and technically by the Bloomberg Philanthropies Data for Health Initiative and partners at the University of Melbourne, Vital Strategies, US Centers for Disease Control and Prevention, National Center for Health Statistics, and the University of Basel Swiss Tropical and Public Health Institute.

Executive Summary

Many low-income countries are considering introducing verbal autopsy (VA) as an integral part of their civil registration and vital statistics (CRVS) systems in order to generate population level cause of death statistics in those parts of the country where there is currently no possibility for medically certified cause of death assignment. There are presently at least 13 countries[1] implementing VA whereby these countries are establishing their technical, process and systems integration needs prior to launching national-scale VA implementation.

The primary purpose of VA in CRVS is to provide statistical trend data at population (not individual) level on the cause-specific mortality fractions for monitoring major health interventions, universal health coverage and sustainable development goals. Such data do not require a verbal autopsy on every death. A sample of deaths is sufficient. But how large should the sample of deaths be? And how should those deaths be selected to ensure results are representative? This strategic guidance document and associated tool are intended to assist such countries.

What are some of the key principles in this strategy? The most important driving principles behind the VA Sampling Strategy and Tool are:

1) Verbal autopsy is not a substitute for medically certified cause of death. It is intended for use where there is no physician, and for generating population level data on proportions and rates of cause-specific mortality. Therefore this guidance is written for countries where a substantial share of the population experiences mortality outside of health facilities and in the absence of medical attendance at death. The tool factors into its calculations the understanding that VAs will be done primarily on community deaths, i.e. those occurring outside of health facilities, even though some deaths occurring in health facilities may not receive a medically certified cause.

2) Verbal autopsies do not need to be conducted on all deaths, but only on an appropriately large random sample of deaths. It is logistically and operationally inefficient to do random VA sampling on individual deaths. Therefore cluster sampling is recommended whereby the cluster unit needs to be decided. The principle we propose is that the minimum cluster sample unit should be the catchment area of deaths that can be reached by a single trained and equipped VA interviewer. Such geographic areas tend to be of a size in which each interviewer would have a work load of 2 to 4 VAs per month, and tend to be approximately the size of census or CRVS enumeration areas (e.g. population sizes between 2,000 and 20,000). This is the minimum cluster unit size. However, some implementation designs may decide on larger cluster units with larger populations and multiple VA interviewers working across the cluster.

3) Sampling should be driven by careful *a priori* decisions on the levels of disaggregation that will be applied in data analysis. At a minimum, the sample size should be adequate to allow analysis of the leading causes of death separately for males and females, and if possible, for the major age groups of neonates, children and adults.

4) Strategic consideration must be given to further geographic disaggregation of analyses (urban/rural, and sub-national (regional/provincial)), especially in countries with decentralized governance of health and social services.

[1] Bangladesh, Colombia, Ghana, Kenya, Morocco, Myanmar, Papua New Guinea, Philippines, Rwanda, Tanzania, Solomon Islands, Sri Lanka, Zambia.

5) Statistical representation requires not just the correct minimum sample size in terms of number of clusters (and VAs) to address the above analyses, but also the drawing of that sample through a random sample of the cluster units from an appropriately constructed sample frame. This Guidance provides a methodology for both calculating the size, and doing random cluster sampling. Given it is highly likely that at some point countries will wish to have disaggregated analyses at least to state, provincial or regional level, we have designed the calculator to be based on an approach of single-stage stratified random cluster sampling proportional to population size.

Who should use this guidance and tool? This CRVS VA Sampling Strategies Guide and its associated Sample Size Calculator Tool are intended primarily for those responsible for providing high quality mortality data in countries where a decision has been made to use VA as part of the CRVS system. It allows CRVS VA managers in such countries to determine **the number and location of geographic units to be sampled to detect a nationally representative change in cause specific mortality fractions or rates in populations where medical certification of cause of death is not yet feasible**.

When should the Guidance be applied? The Guidance and the Tool are expected to be of value to countries who have concluded the pre-test or pilot phases of their VA implementation and who have established the technical, process, and incremental cost considerations with regard to VA implementation at scale.

What questions does the Guidance address? The Guidance addresses four key issues regarding the implementation of CRVS VA at national scale:

1) What are the key logistical considerations to make with regard to the definition of an operational CRVS VA cluster?

2) What are the key strategic considerations to decide with regard to the level of disaggregation at which analyses will be conducted (sex, age, urban-rural, sub-national administrative, trend period, etc.)?

3) What is the minimum number of sample units (clusters) and number of VAs needed given an acceptable uncertainty range for detecting significant CSMF changes over time? Or alternatively, what is the uncertainty range for detecting significant CSMF changes over time given a number of clusters sampled?

4) How should the required sample clusters be selected from the national sample frame?

How can the Guidance be used? The tool allows CRVS managers to:

1) Determine the required sample size for a national or sub-national system given an acceptable uncertainty range for detecting significant CSMF changes over time.

2) Determine the uncertainty range for detecting significant CSMF changes over time given the current or planned deployment of VA.

How does this Guidance help? The relationship between the number of VAs conducted and the resulting uncertainty range for detecting significant CSMF changes over time of various levels of CSMF is not intuitively easy to appreciate. For example, a given number of VAs conducted in a relatively small number of very large clusters will give wider uncertainty ranges compared with those conducted in a larger number of smaller clusters. This has operational and cost implications. Hence, this tool should be used in concert with the CRVS VA Costing Tool. Using both tools will be helpful in making key decisions with regard to strategies for scaling up CRVS VA in national systems.

The tools are available from the Bloomberg Philanthropies Data for Health Initiative CRVS Knowledge Gateway at the University of Melbourne (https://crvsgateway.info) and other CRVS resource portals.

Part A - Principles and Strategy

1. Introduction

This package of guidance materials (Parts A, B, Annexes and the associated CRVS VA Sample Size Calculator Tool) is intended to assist countries with their Civil Registration and Vital Statistics (CRVS)-Verbal Autopsy (VA) scale up and rollout planning. Users will include those tasked with designing and managing the national CRVS VA system, supported by a governing body such as a National Mortality Committee of the National CRVS Committee.

In addition to discussing strategies and principles, this guidance package proposes approaches to conventional cluster sampling methods and provides the statistical rationale, logic, mathematical formulations, and a worked example, for: i) calculating the required number of clusters needed in a VA cluster sample design; and ii) drawing the needed clusters from a national sampling frame.

Every country will have different implementation circumstances and variations in approach. Therefore, this document is necessarily generic in hopes that the considerations, options, and methods provided can be adapted and adjusted to most scenarios where representativeness of VA data is sought.

1.1 Pathways to scale for CRVS Verbal Autopsy

This guidance package is designed to be used after the pre-test or pilot phases of VA implementation during which the technical methods, processes, systems integration, and costing of the CRVS VA are established, and before demonstration and scale-up (See Table 1.1.1).

Table 1.1.1. Pathways to Scale: Phases of CRVS VA Implementation

Phase	Purpose	Example	Scale
Pre-Test	For technical issues	Adapting and testing technologies, instruments, translations, etc.	~ 100 VAs Local scale
Pilot	For process issues	Developing training, supervision, communications, IT processes, initial costing, and SOPs.	~ 1,000 VAs District scale
Demonstration	For systems integration issues	Developing integration with CRVS and HMIS information systems and conducting full costing and sampling strategies.	>1,000 VAs Regional scale emulating proposed national scenarios
Scaling up	For institutionalization	Rolling out to national or sub-national implementation	National sample scale

1.2 Rationale for CRVS VA sampling

In many low-income countries, the majority of deaths occur in areas where there is no medically qualified professional to legally determine and certify the cause of death. Yet such information is critically necessary for effective Civil Registration and Vital Statistics (CRVS) systems and for use in policy and planning. In such countries, there is growing interest to integrate verbal autopsy (VA) into national CRVS systems to cover those deaths currently unreached by medically qualified death certification (de Savigny, Riley et al. 2017). The purpose of such innovations in CRVS is to provide more complete estimates of the cause of death patterns at population level, and not medical certification of individual causes. To achieve this goal, not every death requires a VA; a representative sample of non-medically certified deaths will suffice.

One rationale for adopting a sampling approach is that VA is a sensitive data collection enterprise that requires a household visit and rapport with deceased family members. As such, it is a relatively costly and logistically complex endeavour. Depending upon the numbers of deaths occurring outside of health facilities, costs and complexities could be prohibitive should a country aim to assign a cause to every such death using VA in the national population. A sampling approach, in most cases, can provide statistically valid estimates of the main output of a VA system: Cause-Specific Mortality Fractions (CSMFs).

Therefore, a key question for national CRVS system managers is how many verbal autopsies are needed per year and from which locations, to give sufficiently precise estimates of the cause-specific mortality fractions and rates[2] necessary for reliably informing policy and program decisions?

To date most applications of VA have been in longitudinal health observatory settings such as sentinel Health and Demographic Surveillance Sites (HDSS) or in mortality surveillance systems such as Sample Registration Systems (SRS) or Sample Vital Events with VA (SAVVY) systems (de Savigny, Renggli et al. 2017). The former HDSS sites conduct VAs on the total population in their sentinel sites. Hence, sampling is not an issue. National scale SRS and SAVVY systems are few and each has taken a different approach to establishing their sample size and sampling frame. These are reviewed and reported separately (See papers being produced by Tanzania, Malawi, Mozambique, Zambia, Indonesia, India, and China).

The more specific question of how to sample VAs for integration within CRVS systems has never been fully explored in ways that balance logistical, epidemiological, and statistical considerations related to the specific purposes and needs of CRVS systems. The factors to consider in providing guidance on how to sample for VA are numerous. For example, there are questions with regard to the rapidly changing crude death rates and changing distributions of causes. In highly populous countries, there can be wide geographic and epidemiologic heterogeneity within the country. In addition, in all countries there is socio-economic heterogeneity. Further, political and administrative realities frequently intrude and sometimes over-ride technical and scientific criteria in the sample selection process.

1.3 Rationale for cluster sampling

Given that a VA is not required for every non-medically certified death for population level CSMF estimates, combined with the labour and cost-intensiveness of VA means that the logistical and cost considerations of CRVS VA are most easily addressed by a cluster sampling approach. Cluster sampling brings with it an attendant design effect on the sample size and subsequent analytic considerations. Most cluster sampling methodologies are intended for use in sample surveys, case-control studies or randomized controlled trials of interventions in

[2] This Guidance focuses on the cause specific mortality fraction since this is the prime purpose of verbal autopsy; however the uncertainty ranges for detecting significant CSMF changes as predicted by the CRVS VA Sample Size Calculator Tool are very similar for the corresponding mortality rate of that particular cause of the CSMF.

comparative population groups. Although there are cluster sample methods for studying intervention exposure or health behaviours (e.g. Expanding Programme for Immunization, Demographic and Health Surveys), there are no standard cluster sampling methods for the monitoring of population level cause-specific mortality. This is essentially the challenge posed by CRVS VA.

Sample designs using cluster sampling to control costs and enhance logistical feasibility encounter a design effect due to the fact that variance within clusters is less than the variance in a simple random sample of the whole population. If the design effect is not accounted for, confidence intervals will be incorrectly too narrow and analysts will risk making type 1 errors of concluding differences to be significant when they are not. This design effect can be kept as low as possible by following these general principles:

- Using as many clusters as feasible and affordable;
- Using the smallest operational cluster in terms of population as feasible;
- Using a more constant cluster size rather than highly variable size.

In other words, a design using more small clusters is preferable compared to one based on fewer large clusters.

2. Key principles for National CRVS VA Sampling

There are a number of key principles to consider in embarking on the establishment of a nationally representative CRVS VA system. These are summarized below.

2.1 Non-competition with medical certification

VA is an imperfect tool, but it is generally agreed that it is the best available option for understanding mortality causes in situations where there is no physician in attendance to document the cause of death (D'Ambruoso, Boerma et al. 2016). As a first principle of this Guidance it cannot be over-emphasized that the implementation of CRVS VA must never replace medical certification of cause of death (MCCD) or retard improvements in MCCD. Complete coverage of death registration, death certification, and medical certification of cause for all deaths is the primary goal of the mortality assessment function of CRVS. VA should only be considered where there is no doctor and where there is no possibility to obtain a MCCD, or in implementations strategies where physicians are given the option of using VA to obtain further information to assess the potential cause of death.

The primary mortality objective of CRVS is to produce nationally representative mortality statistics from high quality cause of death data sources annually, disaggregated by age and sex. CRVS needs ways to incorporate information from multiple data sources. For some low resource settings or low CRVS performance countries, such data may only be sourced from VAs for the majority of deaths in the near future. However, the primary aim is to move countries towards greater production and use of high quality, high coverage MCCD data.

In low-income countries, physician certified MCCD data is mainly available from hospitals. While there are usually enough facility deaths to calculate CSMFs as well as cause-specific mortality rates, these estimates are highly confounded by selection biases. CRVS VA from sample systems will never provide as many deaths as seen in hospitals, but VA sample sizes should be chosen to be of sufficient size to provide representative CSMFs at least disaggregated by sex. Secondary mortality objectives of CRVS may be to produce nationally representative trend data on mortality and national and sub-national estimates with further disaggregation, including sex, age, urban/rural, and other dimensions of interest, based on high quality cause of death data sources as frequently as is feasible.

Thus, the availability of CRVS VA will complement the mortality data obtained from hospitals and other health facilities. This poses challenges concerning the analysis and use of mortality data[3].

2.2 The need for effective universal death notification and registration

In any CRVS VA implementation, an active, effective, and universal death notification system is essential. Such a system should notify all deaths in the community to the CRVS system, recording the fact of death with name, date of death, sex, age and essential identity information sufficient to register the death and contribute to CRVS. Nevertheless, given that most CRVS systems are still passive, relying on the family to declare the death to some authority, notification and registration is problematic. Such systems will not ensure that 100% of deaths are notified and available for VA follow-up. The degree of under-notification will significantly affect the sample size. This is accommodated in the CRVS VA Sample Size Calculator Tool but does require some assessment of the under-reporting rate from the pre-test or pilot phases of the implementation.

2.3 Deaths without medical certification of cause of death

Based on the above principle, CRVS VA designs focus predominantly on community (out of hospital) deaths where there is no physician or chance for MCCD. This affects sample size calculations and require decisions on possible CRVS VA deployment options. These considerations concern whether to do VAs on:

- All deaths in the sampled cluster
- Only community (out of facility) deaths in the sampled cluster
- Only deaths without an MCCD in the sampled cluster

Each country will need to weigh the pros and cons for their particular circumstance, knowing the coverage and quality of their facility based MCCD implementation and the cost of VA in the scenario. At current quality levels in many low-income countries, a significant fraction of the MCCD deaths are poorly certified. Even if correctly certified, MCCD deaths may not be coded to appropriate underlying causes. If it is decided that VA should only be done on deaths without an MCCD then excluding MCCD deaths will increase the number of clusters needed for the National CRVS-VA system and increase the cost of VA in countries where a large proportion of deaths have an MCCD. This is even so in Africa, where usually only 30% of deaths are in a health facility. Countries need to know what proportion of deaths currently have an MCCD in order to use the CRVS VA Sample Size Calculator Tool. If that is not known, then they will need to use the proportion of deaths that occur in hospital as a proxy for MCCD coverage. Additionally, this approach will bias VA results to those who are far from physician services. This adds extra analytic complexity and requires weighting estimations when combining data analyses from MCCD hospital enriched data with VA community enriched data.

[3] This Guidance does not discuss VA analytic issues that will be the subject of a separate guidance document from the Bloomberg Philanthropies Data for Health Initiative.

3. Strategic operational considerations for National CRVS VA Sampling

This section of the guide addresses strategic and operational consideration for CRVS VA sampling.

3.1 Defining the operational cluster

A major cost driver for VA in CRVS is the training, deployment, and supervision of the VA interviewer. VA interviewers are deployed at community level such that they can usually reach their assigned or designated catchment area on foot or on bicycle without the need for higher transport costs. A catchment area like this would typically contain a population in the range of 3,000 to 15,000 people and can represent a potential operational VA cluster sample unit. Where crude death rates are in the range of 5 to 10 per thousand per year, such VA interviewers are likely to live in catchment areas that experience between two to ten deaths per month in households where routine visitation is feasible. Such a workload should not overwhelm any other duties the interviewer might have, and is also not so light that VA skills would be lost over time. In rural settings, such areas are often categorized in censuses at Administration Area Level 3 (e.g. Sub-district or ward)[4].

It is therefore important when using this VA Sample Size Calculator Tool to choose a cluster unit with an appropriate size. We suggest setting the ***cluster unit definition as the catchment population of a single VA enumerator or interviewer team***. Hence, the selection of the cluster unit is driven by the logistical realities of the CRVS VA implementation. This is the first decision to set when planning the sampling strategy. Ideally, the geographic boundaries of operational units will correspond with civil registration jurisdictions.

Note the statistical reality that conducting VAs in a small number of clusters with large populations will lead to a larger uncertainty range for detecting significant CSMF changes over time than conducting the same number of VAs in a larger number of clusters with smaller populations.

3.2 How many Cause Specific Mortality Fractions?

VA methods available for CRVS are able to distinguish up to 64 distinct target causes of death. However at population level, CSMFs are usually over dispersed such that the top 20 causes include about 70% of all deaths. In ranking the top 20 causes from the largest to smallest, the first ranked (largest) cause usually accounts for 10 to 20% of all deaths. The top five causes usually include CSMFs of 5% or higher, the top 10 include causes down to about 2% CSMF and the top 20 down to about 1% of all deaths. Therefore providing estimates on the top 20 causes specifically in males and females should cover most causes of interest to policy makers. Hence the CRVS VA Sample Size Calculator Tool allows setting the desired uncertainty range for detecting significant CSMF changes over time for the 1% CSMF level (approximately the 20th ranked cause). The uncertainty range around larger CSMFs will always be narrower than that for the 1% CSMF.

3.3 Disaggregation of results

The disaggregation of CRVS VA data is another fundamental consideration in the design of a national CRVS VA system. For example, national authorities will likely wish to have valid estimates for males and females; for deaths occurring separately in neonates, children and adults; for urban and rural populations; or for sub-national units such as states, provinces, or regions, or for causes addressed by targeted policies (e.g. road traffic or malaria deaths). Each type of disaggregation carries with it important implications for sample size and design. A process

[4] Administrative Level 0 is the national boundary; Level 1 is the state, region or provincial Boundary; Level 2 is the district boundary; Level 3 is county, ward or sub-district boundary; Level 4 is the village, town or municipal boundary; Level 5 is the hamlet, neighbourhood boundary or enumeration area boundary. There is considerable variation among countries in the use of these terms and not all countries use all levels, in the same way.

should be established whereby the necessary stakeholders are consulted, and trade-offs in terms of cost and complexity are assessed under varying scenarios for disaggregation before a final decision is reached.

This CRVS VA Sampling Strategy Guidance and Calculator Tool will calculate the number of clusters that need to be selected to provide estimates of the CSMFs for the top 20 causes of deaths in a given population. A major driver of the sample size will be the number of discrete populations for which you require these top 20 causes (e.g., disaggregations by sex, age, geographic area, socio-economic status, etc.). It should be borne in mind that the number clusters directly affects the number of VAs and interviewers required to implement a CRVS VA system

3.3.1 Male-Female disaggregation

It is essential in any CRVS system that results are disaggregated by sex. Therefore, if male-female disaggregation is required the CRVS VA Sample Size Calculator Tool will approximately double the sample size in order to provide estimates separately for both male and female causes of death (See Annex A.10.1 andB.7). Thus, the user-configurable settings of the CRVS VA Sample Size Calculator Tool will allow the choice to provide CSMFs for the top 20 causes in males and females at national level.

3.3.2 Age Group disaggregation

Modern VA questionnaires and diagnostic algorithms are designed to provide age group specific results for neonates (zero to 27 completed days of age), children (28 completed days to 11 years of age); and adults (12 years and above). These age groups constitute very different proportions of the denominator population and of course very different proportions of the total population mortality (see Table 3.3.1)

Table 3.3.1. Approximate VA age group shares in low-income countries.

Verbal Autopsy Age Group	Share of total population	Share of total mortality	Number of VA detectable target causes
Neonates (0-27 days of age)	~3%	~5%	6 to 7
Children (28 days to 11 years of age)	~23%	~15%	26 to 57
Adults (12 years and older)	~74%	~80%	26 to 57

The Sample Size Calculator Tool is designed to estimate the needed sample size for the whole male and female population of all ages. Some countries may want a particular focus on a smaller age group such as neonates or children. For neonates, who represent a very small proportion of the total population, a larger sample size would be needed. However there is less need to disaggregate by sex in neonatal mortality analysis, and there are relatively few causes of neonatal mortality detectable by verbal autopsy. Neonatal causes tend to be more equally distributed, so that uncertainty ranges can be estimated around 10% CSMF rather than 1%. For a focus on the child age group by sex, one would need to approximately quadruple the sample size to obtain the same uncertainty range across the CSMFs as found for the total (male/female) estimation.

3.3.3 Urban-rural disaggregation

Another desirable disaggregation may be by urban and rural populations. With the sampling strategy proposed in this guidance, the urban-rural ratio is self-weighting, and the sampling strategy will allow national estimates of

CSMFs, accounting for urban rural status. However, if separate estimates of urban and rural CSMFs are needed, then the Sample Size Calculator Tool can be used to independently calculate and draw sample clusters from urban and rural clusters separately.

3.3.4 Sub-national administrative disaggregation

Politically, some countries may wish to provide sub-national estimates of CSMFs based on administrative levels, most likely down to Administrative Level 1 (e.g. Province, Region, or State level). This will increase the sample size required, along with the costs and scale of CRVS VA. However, if at the sub-national level, estimates with sufficient statistical power are only required on the top five causes or so, a national-level sample may be adequate.

3.4 De-duplication

If VA's are purposely only done on community deaths or on deaths that do not have an MCCD, there will be instances where VAs will still be done on deaths having had an MCCD. Community key-informants notifying deaths to the VA system may not know the MCCD status of the death, and indeed the VA respondent at the household level may not know the status, or understand the difference between a death certificate and a medical certificate of cause of death or even a burial permit. In such instances, there is a need to avoid the risk of double counting and double registration through proper system design and the use of unique identifiers. Addressing duplicate registration should not be solely a Ministry of Health or Civil Registrar's data management problem but rather avoided in the design of processes. De-duplication is most efficiently done manually in combination with IT solutions as part of system design.

4. Considerations for framing the CRVS VA sample

4.1 The Sample Frame

To calculate the sample size and determine the operational clusters for CRVS VA, a sampling frame needs to be created. This is a complete inventory of all clusters eligible to be sampled in the country. Part B of this Guidance Package provides a sample template for such a database (Preparing the sampling frame). Establishing a sample frame requires the compilation of available administrative and demographic information for each cluster included:

- name of the cluster;
- a unique administrative ID or census code;
- its parent administrative hierarchy of Region and District;
- current estimated population;
- estimated projection of the current crude death rate;
- derived estimate of the number of expected deaths per year.

Potential governmental sources of the above information are suggested in Part B of the manual.

Some additional cluster information might be valuable. This could include:

- urban or rural status;
- area in km^2;
- derived population density per km^2;
- the presence of hospitals and health facilities per cluster; or
- the number of lower administrative area units (e.g. number of villages).

- clusters that might host unusual populations (such as national parks or reserves, refugee camps or nomadic populations) may also be marked and considered in inclusion/exclusion decisions.

4.2 Inclusions and exclusions

It is useful to open this sample frame database in a Geographic Information System (GIS) or any other GPS application that hosts the GIS shape or boundary files for the clusters, assuming the clusters have the correct name or index field. This allows visualizing the size and spatial distribution of the clusters on a national or regional map. The GIS mapping of the clusters will make evident certain extremes, such as very large area clusters with very sparse population densities, or very small, densely populated clusters in cities or slums. The presence of refugee camps, national parks and reserves, and other non-representative populations should also be mapped. You may wish to exclude such areas, including very low population density clusters from your sample frame. Such maps are also useful in understanding and communicating how the final sample will be distributed across the country.

5. Considerations for calculating the CRVS VA sample size

This section of the Guidance Package gives an overview of the CRVS VA Sample Size Calculator Tool and summarizes what input parameters it requires, and what outputs it delivers. A Step-by-Step Manual is provided in Part B of this Guide along with Annexes that provide the details of the statistical calculations it uses. Part B also provides a real-world worked example of how the Tool can be applied in determining a national CRVS VA sample size.

This guidance assumes that the national stakeholders have made decisions concerning the operational cluster unit definition and the required disaggregation level of the results as described in Section 3.

5.1 Statistical approach used in the CRVS VA Sample Size Calculator Tool

The statistical approaches used in this CRVS VA Sample Size Calculator Tool are based on the guidance given by Hayes and Moulton in their book on cluster randomized trials (Hayes and Moulton 2009). Additionally, further literature was consulted to take into account less trial specific setting as they are present in CRVS VA systems. A more detailed description of the statistical basis of the CRVS VA sample size calculations and all formulae are elaborated in Annex A. Here we will focus on the most important parameters a user needs to understand in order to make use of the Sample Size Calculator Tool.

5.2 Overview of the CRVS VA Sample Size Calculator Tool

The CRVS VA Sample Size Calculator Tool is a complex statistical calculator currently in MS Excel, but with a very simple user interface. User defined inputs are entered into yellow shaded cells. At start up, the user selects the country name from a drop-down list and enters the year on which the input data is based on. There is an option to indicate if the sample is for a subnational stratum such as a Region.

The tool works in two modes:

1) **Mode 1** calculates the required sample size (number of clusters) for a pre-determined acceptable uncertainty range for detecting significant CSMF changes over time for the CSMF of 1% which must be entered in a designated cell.
   Or
2) **Mode 2** calculates the uncertainty ranges for detecting significant CSMF changes over time for a given number of clusters which must be entered in a designated cell.

Finally the analyst indicates (a) whether or not there is an intention to disaggregate results by sex with selecting yes or no and (b) if the population size of each eligible cluster is available. (See Figure 5.2.1)

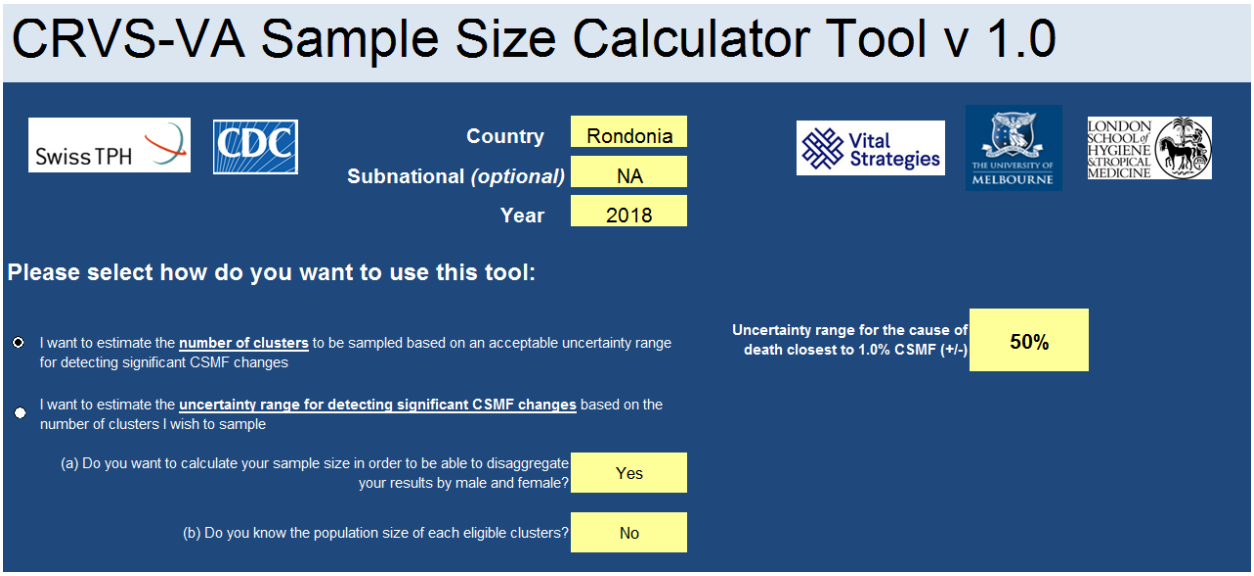


Figure 5.2.1 Example of the start-up screen from the CRVS VA Sample Size Calculator Tool

5.3 Tool input parameters required

If the population size of each eligible cluster is not known the tool requires one mode specific input (5.3.1 or 5.3.2) and five general inputs to calculate the above described scenarios (5.3.3-5.3.7) (See Figure 5.3.1). For the scenario where the population size of each eligible cluster is known, see 5.3.8.

5.3.1 Number of clusters

To estimate the uncertainty ranges for detecting significant CSMF changes over time based on a given number of clusters, you will need to enter the number of clusters you are expecting to sample.

5.3.2 Maximum acceptable uncertainty range

The uncertainty range is the range within which changes in CSMFs over time will not be detectable with sufficient statistical power. Note: the range is given in percentage change, which is not to be mistaken with percentage point change. We have indexed the Calculator to allow you to specify the acceptable uncertainty range for detecting significant CSMF changes for the smallest CSMF of interest (causes in the range of 1% of total deaths) as for this cause the range will be the widest. Uncertainty ranges at all higher CSMFs will always be narrower. Thus, the value to be entered is the uncertainty range (percentage change) of the 1% CSMF that is acceptable. For example, setting the acceptable uncertainty at 50% (as in Figure 5.2.1) means that the tool will calculate a sample size capable to detect a change in CSMF over time of 50% or more, meaning for the 1% CSMF an increase to 1.5% or a decrease to 0.5% in the next time period.

5.3.3 Mean cluster population

The Calculator needs to know the mean cluster population in order to estimate the numbers of deaths, and hence VAs needed. Ideally the cluster definition should be chosen such that the range of cluster populations is not too large.

5.3.4 Crude death rate

The annual crude death rate (CDR) per 1,000 is a key input parameter that affects CRVS VA sample size. You will need a national estimate. This might be available from a recent census, but note that in many low-income countries, CDR changes measurably from year to year. In most low-income countries, the CDR is presently falling. Thus, for the purposes of estimating the sample size for a national CRVS VA system it is better to be conservative and choose a lower expected value for CDR. Using a lower CDR increases the likelihood that enough clusters are sampled to produce the desired estimates. CDRs can be found online from a number of sources that model and forecast CDR trends for each country (See Part B and Annex C.1).

5.3.5 Number of years to aggregate for trend

The Calculator will need to know how many years you aim to aggregate for the trend analysis. As a default it adjusts for a three-year follow up period. This means that it is assumed that deaths occurring within the three years will be aggregated and CSMFs from the first three years will be compared with CSMFs of the second three years. We make the assumption that there is no re-selection of clusters after the first three years. The calculator estimates the number of clusters needed to detect a significant change in one CSMF between the first and the second three years given a pre-determined uncertainty range for detecting significant CSMF changes.

5.3.6 Percentage of deaths with MCCD

Assuming a strategic design decision has been made that VA will be conducted only on deaths that occur outside of health facilities or for deaths without a MCCD (see Section 2.2) a national estimate of the proportion of deaths in health facilities or with an MCCD will be required, respectively.

5.3.7 Under-notification and non-response rate

It is very likely that it will not be possible to conduct a VA on every eligible death in the cluster. One problem could be that deaths for which VA should be done are not notified. This is particularly of concern with neonatal deaths. It could also be that the death is notified, but due to implementation challenges (e.g. household not reachable, family moved away) the VA interview is not conducted. Further, there is the possibility that the family of the deceased refuses to consent to the interview. This is what we call the "*Under-notification and Non-response Rate*" in the tool. An estimate of the proportion of such missed deaths must be made as this rate strongly influences the sample size. It is crucial to be conservative in this estimate. If this rate is under estimated, the sample size might be too small to draw conclusions with acceptable uncertainty ranges. From experience so far using VA in CRVS, the under-notification and non-response rate has been as high as 40%. Even with solving most implementation problems, under-reporting will likely not be less than 10%. A careful analysis of results from the pre-test or pilot phases should be conducted to estimate this CRVS VA performance parameter realistically (See Table 1.1.1).

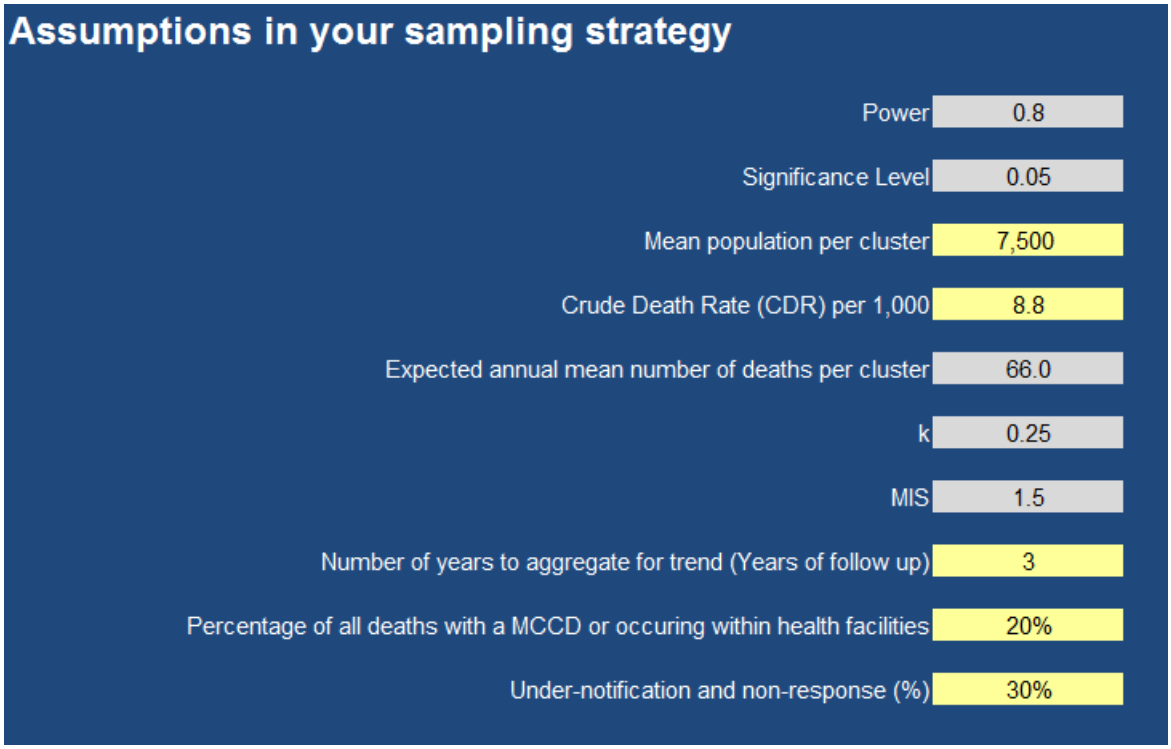


Figure 5.3.1. Example of the input parameters screen from the CRVS VA Sample Size Calculator Tool if population size of eligible clusters is not known

5.3.8 Scenario where population size of eligible clusters is known

If the population size of each eligible cluster is known the tool requires the analyst to list all clusters with their population sizes as well as cluster specific CDRs (Figure 5.3.2). In case the latter is not available the national CDR (5.3.4) can be listed for all clusters. Note: Eligible cluster are all clusters in your final sampling frame after having excluded not eligible clusters. The tool will calculate the expected annual number of deaths in each cluster and the annual mean number of deaths per cluster (using the harmonic mean). The inputs for "Mean cluster population" (5.3.3) and "Crude Death Rate" (5.3.4) will subsequently not be needed anymore (Figure 5.3.3).

Name of the cluster unit	Cluster population size	Crude Death Rate (CDR) per 1,000	Expected annual number of deaths
Cluster 1	6,333	9.12	58
Cluster 2	11,037	6.55	72
Cluster 3	19,090	6.22	119
Cluster 4	23,158	7.70	178
Cluster 5	7,735	6.28	49
Cluster 6	21,454	5.00	107
Cluster 7	5,661	6.89	39
Cluster 8	9,096	6.69	61
Cluster 9	5,476	6.89	38
Cluster 10	15,851	8.45	134
Cluster 11	27,640	6.08	168

Figure 5.3.2 Example of the list with all eligible clusters and their corresponding population size and CDR

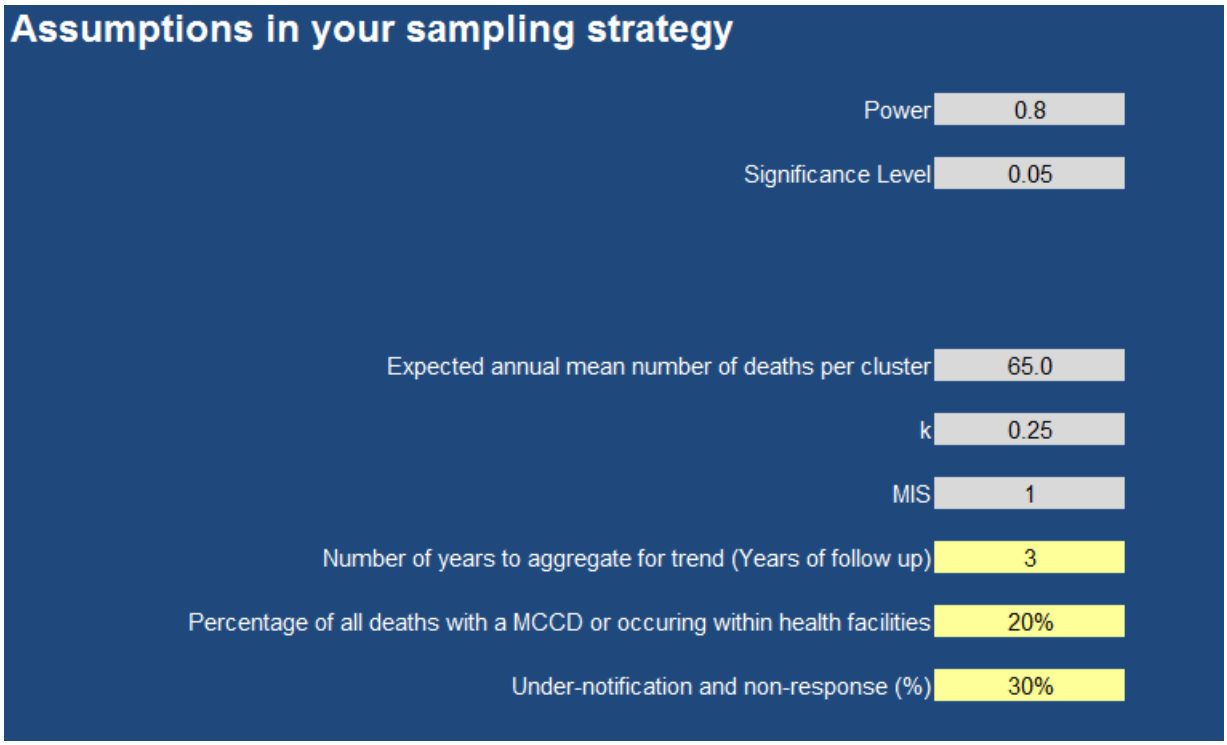


Figure 5.3.3 Example of the input parameters screen from the CRVS VA Sample Size Calculator Tool if population size of eligible clusters is known

See Part B and the Annex A for the explanation of the pre-set values for power, significance level and design related parameters (k, MIS), and how to change these if desired.

5.4 Tool output parameters produced

The CRVS VA Sample Size Tool produces five outputs instantly once the inputs are entered (See Figure 5.4.1).

5.4.1 Number of clusters required

The main output of the Calculator is the number of clusters required for the scenario inputs. This determines all other outputs listed, and provides the basis for costing the scenarios.

5.4.2 Estimated total population in the sample

It may be of interest to know the sampled population size in aggregate across the sampled clusters in order to understand what percent of the total national or stratum population is under VA surveillance.

5.4.3 Estimated number of deaths in the sample per year

This output is the number of expected deaths per year in the total sample of clusters. It should not be confused with the total number of VAs required since this will also include deaths with a MCCD, occurring inside the health facility, or missed due to under-notification and non-response.

5.4.4 Estimated number of VAs needed per year

This is the number of VAs needed per year and a useful output if the incremental cost of each VA is known.

5.4.5 CSMF uncertainty ranges

Finally a table of CSMF uncertainty ranges (Uncertainty as % of CSMF, Lower and Upper bound of uncertainty range) for detecting significant CSMFs changes for ten major levels of CSMF from 1% to 25% is given. In other words the lower and upper bound indicate the lowest detectable difference between the CSMFs from the first three years and the second three years. This means the CSMF of the second three years must be equal or outside these boundaries in order for the change to be statistically significant. The CSMF levels span the full range of CSMFs expected for the causes of major public health importance. The range indicated for the 1% CSMF should equal the range entered in the input parameters. The uncertainty range at higher CSMFs will all be narrower. It does not matter what the specific cause is for any given CSMF, the uncertainty range is relative to the CSMF level. See Figure 5.4.1 for all outputs specified by the inputs in Figure 5.3.1.

Results

Number of clusters required	140
Estimated total population in the sample	1,050,000
Estimated number of deaths in the sample per year	9,240
Estimated number of VAs needed per year	5,175

CSMF UNCERTAINTY RANGES			
CSMF level	Uncertainty (% of CSMF)	Lower bound	Upper bound
25.0%	17%	20.8%	29.2%
20.0%	17%	16.5%	23.5%
15.0%	19%	12.2%	17.8%
12.5%	20%	10.1%	14.9%
10.0%	21%	7.9%	12.1%
7.5%	23%	5.8%	9.2%
5.0%	26%	3.7%	6.3%
3.0%	32%	2.0%	4.0%
2.0%	38%	1.2%	2.8%
1.0%	50%	0.5%	1.5%

Figure 5.4.1. Example of the output parameters screen from the CRVS VA Sample Size Calculator Tool

6. Considerations for selecting the CRVS VA sample clusters

This section of Part A outlines key operational considerations for the **selection of the required number of clusters** once the sample size is calculated as above. A manual of the full step-by-step details for how to draw the sample is provided in Part B section 4.

The key principles of the sample selection applied in this guidance package include:

- Using a suitable sample size
- Using a **sampling frame** which is a complete list of all cluster units eligible to be sampled
- Using the most simple sample selection strategy possible
- Drawing and implementing the sample exactly as designed
- Providing good sampling method documentation

6.1 Defining the sample selection strategy

This section describes the selection strategy of the CRVS VA sample from the sampling frame once the cluster sample size is known and the sample frame of eligible clusters has been prepared. To foreshadow the intention, we recommend ***"stratified single-stage proportional to population size cluster sampling"*** and summarize the background and logic for this below.

6.1.1 Stratification

Simple random sampling is one option to select the required number of clusters from the sampling frame. However this can cause problems; the resulting sample may not adequately reflect characteristics of the country.

For example, in a simple random sample, clusters from rural areas could be under-represented or over-represented compared to the proportion of the population living in such areas. Stratified sampling tries to address this problem by using information about cluster characteristics to choose a more representative sample. Therefore, it is recommended to stratify your sample, if possible. For example, it might be desirable to ensure that clusters are representative for the populations within each major administrative zone (e.g. Level 1, Region, State, or Province). This is the most common sub-national level of disaggregation that countries are interested in, both politically and epidemiologically.

Independent of the decision to stratify or not, each sample cluster will need to be randomly selected from the sampling frame. Here we present three options for doing this. For further details on this, also see Part B, section 4.1 and the worked example in Annex B.

6.1.2 Simple random sampling

In simple random sampling, all potential sampling units (clusters) are known and can be listed, and each unit (cluster) within the sampling frame has an equal probability of being selected. The selection would be done by creating random numbers between 1 and the number of clusters in the sampling frame. The number of random numbers to be created is given by the number of required clusters.

6.1.3 Systematic sampling

In systematic sampling, the sampling frame is sorted according a selected variable (e.g. alphabetic order of administration level 1 names). In a next step, the number of clusters in the sampling frame (or, in the case of stratified sampling, in the stratum) is divided by the number of clusters to be sampled (Bierrenbach 2008). This reveals the sampling interval. Thirdly, a random number between 1 and the sampling interval is chosen as the random starting point in the sampling frame. The cluster at the position of the random starting point in the sorted sampling frame is the first cluster to be selected. The other clusters are then selected at regular intervals according to the size of the sampling interval.

6.1.4 Probability Proportional to Size sampling

The main problem of the two options described above is that the sampling does not result in the most representative selection of clusters. For example, you might end up selecting clusters with low populations. To overcome this problem, we can use Probability Proportional to Size (PPS) sampling.

PPS sampling is a sampling procedure under which the probability of a cluster being selected is proportional to the population size of that cluster. Given that the number of deaths will be generally correlated with the population size, the sample will be self-weighting.

6.1.5 Stratified single-stage cluster PPS sampling

To address the above concerns we propose stratified single-stage cluster PPS sampling. Note however that sampling can be done in one stage (single-stage) or in multiple stages (multi-stage). In single-stage sampling, we sample only once although you could use several different sampling strategies in combination. For example, you could stratify your sample first by regions and then sample clusters from many within a stratum (actual sampling by simple random sampling or PPS sampling). On balance, we recommend using a ***"stratified single-stage PPS***

cluster sampling" approach for VA implementation if possible. The associated CRVS Sample Size Calculator Tool is designed for such kind of sampling approach[5]. See Part B Section 4.2 for details on how to apply this approach.

6.2 Documenting the sample method

In preparing this guidance package and consulting a number of countries that operated sample registration systems we found that corporate memory on how the sample size was calculated and drawn had been lost over the years. Based on the CRVS Sample Size Calculator Tool inputs, one function of the Tool that we have added is a function to produce not only tabular outputs as shown in Figure 5.4.1 but also a narrative text that can be saved and used in documentation, reports or publications to summarize the essential information needed to communicate and archive the sample calculations. Below is an example of the sample documentation script. The tool automatically adds the input and output values indicated into the script output.

This CRVS VA sample size was calculated assuming an under-notification and non-response rate of a%, a MCCD/health facility death coverage of b%, an annual crude death rate of c per 1,000, and a mean population per cluster of d. The sample size was designed to give e% power of obtaining a significant difference (p<f) for a change of g% in the cause specific mortality fraction of 1% when monitored over h years. Design related parameters were assumed to be k_m = i and MIS = j. The sample size was doubled in order to have disaggregated results for male and female. Based on these inputs, the estimated sample size required was k clusters (and l verbal autopsies per year). The estimate is based on the cluster sample size formula for proportions from Hayes and Bennett for matched studies (Hayes and Bennett 1999). The sample is intended to be drawn based on stratified single-stage cluster proportional to population size sampling.

In case the population size of each eligible cluster is known the first sentence would be modified to:

This CRVS VA sample size was calculated assuming an under-notification and non-response rate of a%, a MCCD/health facility death coverage of b%, and an expected annual mean (harmonic mean) number of deaths per cluster of d.

7. Limitations

There are certain limitations of the sample size calculations and the sample selection strategy, some of which are discussed here.

For the sample size calculations, we considered all input parameters (see section 5.3) to be relatively constant over time. If they were to change significantly over time this could affect the sample size calculations:

- **CDR:** This could be a problem if the CDR decreases significantly because then the number of required clusters would be underestimated. Thus, we recommend in section 5.3.4to choose a lower expected CDR value in order for the sample size calculation to be conservative.

[5] An alternative to single-stage sampling would be to do multi-stage sampling. This would entail sampling a few regions from many (first stage sampling), and then selecting a pre-determined number of clusters within the regions sampled from the first stage (second stage sampling). However, such a sampling strategy will have implications on the sample size calculations. It is likely to increase the design effect[5] and thus lead to a larger required number of clusters to preserve the level of statistical power desired. Alternatively, if the number of clusters is fixed, the uncertainty ranges for detecting significant CSMF changes will increase.

- **Cluster size:** Assuming a constant k, an increase in cluster size would reduce the number of clusters needed and a decrease in cluster size would increase the number of clusters needed.

For the sample selection strategy, we have not discussed a situation where some clusters would be purposively sampled based, for example, on political considerations. Nor have we treated capital or large cities any differently in terms of sampling and the number of clusters selected. Depending on circumstances, however, such non-statistical sampling of some clusters may be a practical necessity.

8. Scaling up strategy and follow-up period

Moving from the pre-test, through the pilot and demonstration phases (Table 1.1.1) whereby the methods, processes, systems integration and costing of CRVS VA are established, it is likely that reaching the minimum sample size needed for the national level male and female CSMFs will take a year or more. It will then take some years to begin to see detectable trends – particularly for less common causes. The Sample Size Calculator Tool addresses this problem by allowing to aggregate deaths across a user specified number of consecutive years (see section 5.3.5). We make the assumption that there is no re-selection of clusters after the first set of consecutive years of follow up, and that the CRVS VA will continue for some years in the same clusters, given the logistical and cost implications of equipping training and supervising VA activities. As the outcome measure is mortality, we do not expect any detectable Hawthorne effect of long-term monitoring of mortality in the same clusters (Ref from INDEPTH, Matlab, etc.). During the initial scaling up to national level, it is unlikely that countries will achieve the sample size sufficient for sub-national analysis although for example disaggregating by urban and rural could be feasible. Scaling up over time to permit sub-national or other sub-group analyses would involve adding more clusters, rather than selecting different clusters.

9. Conclusions

Who is this Guidance for? This CRVS VA Sampling Strategies Guide and its associated Calculator Tool are intended primarily for those responsible for providing high quality mortality data in countries where a decision has been made to use VA as part of the CRVS system. It allows CRVS VA managers in such countries to determine **the number and location of geographic units to be sampled to detect a nationally representative change in CSMFs or rates in populations where medical certification of cause of death is not yet feasible**.

When should it be applied? The Guidance and the Tool are expected to be of value to countries who have concluded the pre-test or pilot phases of their VA implementation and who have established the technical, process, and incremental cost considerations with regard to VA implementation at scale.

What questions does it answer? The Guidance addresses three key issues regarding the implementation of CRVS VA at national scale:

1) What are the key logistical considerations to make with regard to the definition of an operational CRVS VA cluster?
2) What are the key strategic considerations to decide with regard to the level of disaggregation at which analyses will be conducted (sex, age, urban-rural, sub-national administrative, trend period, etc.)?
3) What is the necessary number of sample units (clusters) and number of VAs needed given an acceptable uncertainty range for detecting significant CSMF changes over time? Or alternatively, what is the uncertainty range for detecting significant CSMF changes over time given a number of clusters sampled?
4) How should the required sample clusters be selected from the national sample frame?

How can it be used? The tool allows CRVS managers to:

1) Determine the required sample size for a national or sub-national system given an acceptable uncertainty range for detecting significant CSMF changes over time.
2) Determine the uncertainty range for detecting significant CSMF changes over time given the current or planned deployment of VA.

How does this help? The relationship between the number of VAs conducted and the resulting uncertainty range of various levels of CSMF is not intuitively easy to appreciate. For example, a given number of VAs conducted in a relatively small number of very large clusters will give wider uncertainty ranges for detecting significant CSMFs change over time compared with those conducted in a larger number of smaller clusters. This has operational and cost implications. Hence, this tool should be used in concert with the CRVS VA Costing Tool. Using both tools will be helpful in making key decisions with regard to scaling up CRVS VA in national systems. The tools are available from the Bloomberg Philanthropies Data for Health Initiative CRVS Knowledge Gateway at the University of Melbourne (https://crvsgateway.info) and other CRVS resource portals.

Sampling Strategies for National Scale CRVS Verbal Autopsy Planning:

A Guidance Document and Sample Size Calculator Tool

Part B: Methods and Tools

Version 2.4

July 26, 2018

Review Version

Glossary

Cause-specific mortality fraction: The proportion of deaths caused by a specified cause of death relative to the total number of deaths occurring in a given period of time.

Cause-specific mortality rate: A mortality rate that specifies deaths according to their cause (International Epidemiological Association 2014) expressed as the number of people dying due to a specified cause of death relative to the total number of population for a given period, usually denominated by 100,000 population per year.

Cluster sampling: A method of sampling in which the members of a population are arranged in groups (the 'clusters') based on defined cluster units, and a number of clusters are selected at random within which all population members are included.

Cluster sample design: refers to the use of clusters in sample surveys as opposed to simple random sampling.

Cluster size: Is the number of individuals in a single cluster contributing to the denominator of the outcome to be measured across clusters.

Cluster unit: The sampling unit is "the unit into which a sampled population is divided for purposes of selection for study" (International Epidemiological Association 2014). In cluster sampling, these units may be geographic or administrative regions, communities, households, or other aggregates or entities.

Coefficient of variation between clusters: "The ratio of the standard deviation to the mean" (International Epidemiological Association 2014). The coefficient of variation between clusters measures the variation of the true outcome measure between clusters.

Coefficient of variation in cluster size: "The ratio of the standard deviation to the mean" (International Epidemiological Association 2014). The coefficient of variation in cluster size measures the variation of cluster size between clusters.

Complex sampling: In contrast to simple random sampling, sampling which includes stratification, multistage clustering or over sampling.

Crude death rate: "An estimate of the portion of a population that dies during a specified period unadjusted by age. The numerator is the number of persons dying during the period; the denominator is the number in the population, usually estimated as the midyear population" or person-years lived in the year. (International Epidemiological Association 2014) If not stated otherwise within this document the specified period is one year and rates are given per 1,000 population.

Design Effect: An effect of a study design feature on the performance or outcome of a statistical procedure. A specific form is the effect attributable to intra-class correlation in cluster sampling. The design effect (deff) for a cluster design is the ratio of the variance for that design to the variance calculated from a simple random sample of the same size. (International Epidemiological Association 2014).

Implicit stratification: A means of stratifying through geographic sorting of the sample frame, coupled with systematic proportional to size sampling (PPS).

Intra-cluster Correlation Coefficient: "Intra-class correlation in surveys and group-randomized studies is the extent to which members of a group (cluster) resemble each other more than they resemble members of other groups (clusters)." (International Epidemiological Association 2014).

Matched design: Two groups of observations are compared, which originate from matched pairs. Matched pairs is a term used for observations arising from either two individuals or clusters that are individually matched on a

number of variables, for example, age, sex, etc., or where two observations are taken on the same individuals or clusters on two separate occasions (Everitt and Skrondal 2010).

Maximum possible Inflation in sample Size: The additional amount by which the sample size required for a cluster design needs to be multiplied to obtain the size required when cluster sizes are variable rather than equal.

Multistage cluster sampling: First sampling subnational geographic areas, and then sampling clusters.

Uncertainty range: The range within which changes in the outcome measure from one time period to the subsequent time period will not be detectable with sufficient statistical power. The range is given in percentage change.

Medical certification of cause of death: The situation where a licensed physician or another designated health professional determined the causes of death and these causes were documented on a medical certificate of cause of death form (International Epidemiological Association 2014).

Person-years of follow up: "A measurement combining persons and time as the denominator in incidence and mortality rates when, for varying periods, individual subjects are at risk of developing disease or dying. It is the sum of the periods of time at risk for each of the subjects. The most widely used measure is person-years. With this approach, each subject contributes only as many years of observation to the population at risk as the period over which that subject has been observed to be at risk of the disease; a subject observed over 1 year contributes 1 person-year, a subject observed over a 10-year period contributes 10 person-years." (International Epidemiological Association 2014).

Power: "Roughly, the ability of a study to demonstrate an association or effect if one exists. The probability that the test hypothesis will be rejected if it is false; it is equal to 1- β, where β is the probability of type ii error (failing to reject a false null hypothesis)." (International Epidemiological Association 2014).

Primary sampling unit: Geographically-defined administrative unit selected at the first stage of sampling.

Probability sampling: Selection methodology where by each population unit (person, household, cluster, etc.) has a known, non-zero chance of inclusion in the sample.

Probability Proportional to Size Sampling: A sampling procedure under which the probability of a unit (e.g. a cluster) being selected is proportional to the size of the unit.

Sampling frame: "A list of members of the population of interest is called the sampling frame." (Upton and Cook 2014) In cluster sampling this is a list of all clusters eligible to be sampled.

Sample Registration System: A Sample Registration System is one in which the fact of vital events, such as births and deaths, are recorded or registered but without cause of death documentation.

Sample vital events with Verbal Autopsy: An SRS system that includes active follow-up of deaths in the community to determine their likely cause of death is called "Sample Vital Events with Verbal Autopsy (SAVVY)."

Significance level: A pre-specified cut-off or alpha level for declaring a result "statistically significant", typically 0.05 (International Epidemiological Association 2014).

Stratification: "The process of or result of separating a sample into several subsamples according to specified criteria, such as age groups, socioeconomic status, etc." (International Epidemiological Association 2014). In stratified sampling the population is divided into subsets (called strata), usually mutually exclusive subnational geographic areas, within each of which an independent sample is selected.

Systematic sampling: Selection from a list, using a random start and predetermined selection interval, successively applied.

Unmatched design: Two groups of observations are compared, which originate from a completely different set of individuals or clusters.

VA cluster unit definition: The catchment population of a single VA enumerator or interviewer team.

VA target causes: A cause of death which can be identified using a particular VA questionnaire and algorithm combination.

1. Introduction

Part A of the guidance packages described the principles and strategies for national scale CRVS VA sampling, including strategic operational considerations as well as considerations for framing the sample, calculating the sample size and selecting the actual CRVS VA sample clusters. Part B of this guidance package provides a step-by-step manual for how to implement the principles and strategies outlined in part A:

1. how to prepare the national sample frame of all eligible CRVS VA clusters;
2. how to use the CRVS VA Sample Size Calculator Tool to determine the number of clusters to sample for CRVS VA; and
3. how to draw the required number of clusters from the sampling frame.

The Annexes include the statistical basis of the CRVS VA sample size calculations, a worked example of the entire process as well as a number of useful resources.

2. Preparing the sampling frame

To calculate the sample size and determine how to select the number of operational clusters for CRVS VA a sampling frame needs to be created. To do so we start with a complete inventory of all possible clusters in the country entered into a spreadsheet or database. Available demographic information on each cluster should be included in the inventory. This may include the current population of each cluster, its area in km^2, the population density, and urban or rural status and estimated projection of the current crude death rate. This is most easily done by preparing a flat MS Excel table listing all of the cluster units by name or administrative or census code. Each cluster unit listed should be associated with its administrative hierarchy of Region, District, etc. in which it resides and assigned a unique ID. The necessary attributes (listed below) for each cluster in the frame can then be entered as shown in Figure 6.2.1.

	A	B	C	D	E	F	G	H	I	J	K
1	Zone	Region	District	VA operational area [cluster]	Unique ID of the cluster	Estimated population size	Estimated CDR [/1'000]	Estimated number of expected deaths (=F*G/1,000)	Area [sqkm]	Population density [/sqkm] (=F/I)	Urban-rural status [u/r]
2	Zone 1	Region 3	District 21	Cluster 342	1321342	15,000	6.54	98.1	210	71.4	u
3											
4											
5											
6											
7											
8											
9											
10											
11											
12											

Figure 6.2.1 Example of a sampling frame database

Estimated population size

Estimated population sizes for the target year down to Administrative Level 4 should be available from the national bureau of statistics. However if this is not the case, the national bureau of statistics normally provides decadal census data down to Administrative Level 4, with inter-censal growth rates to Administrative Level 1. Use the census population from the census year for each cluster in the sampling frame database. Then add a column to enter the inter-censal growth rate per annum from the Administration Level 1 in which the cluster resides (or any lower Administration Level if available). In a last step estimate the size of the population in the target year using the following formula:

$$ExpPop = Pop_{Census\ Year} * e^{\left(\frac{r}{100} * t\right)}$$

Where r is the inter-censal growth rate per annum and t the number of years between the census year and the year for which the population size ought to be calculated.

Estimated crude death rate

The next column records the estimated CDR for the cluster. To calculate the estimated CDR you will need the national CDR of the target year (see Part A 5.3.4) and the national CDR from the last census, where also CDR rates for subnational levels are available. Pick the lowest administrative level for which CDRs were reported. Finally, calculate the ratio of the subnational CDRs to the national CDR for the year of the census and then apply this ratio to the national CDR of the target year.

Estimated expected number of deaths

Multiplying the cluster CDR times the cluster population divided by 1,000 will provide the expected number of deaths per cluster per year.

Area and population density (optional)

The next column in the database could be the area of the cluster in km^2. This should be available from any national digital cartography office or could be calculated using a Geographic Information System (GIS) and shape files (see below "Visualization of clusters in maps using GIS"). The next column is for the calculated value of the population density in population per km^2. This is useful in understanding potential clusters to exclude. Particularly sparsely populated clusters may pose insuperable logistical and cost challenges which may justify their exclusion from the sample frame.

Urban-Rural status (optional)

The last column in this example of a sampling frame indicates whether a given cluster is predominantly urban or rural.

Additional information

Some additional cluster information might be relevant as well. This could be the presence of hospitals and health facilities per cluster or the number of lower administrative area units (e.g. number of villages). Clusters that might host unusual populations (such as refugee camps or nomadic populations) may also be considered in selection/exclusion decisions.

3. Calculating the sample size

This section of the Guide deals with information that will be needed in order to use the CRVS VA Sample Size Calculator Tool and gives step-by-step instructions on how to actually use the Tool. It assumes that primary stakeholders have made the key decisions concerning the operational cluster unit definition and the required disaggregation level of the results as described in Part A, Section 3.

3.1 Preparatory steps

In order to calculate the needed number of clusters for CRVS VA implementation at national scale or at the selected level for disaggregation, we need to have certain input parameters at hand. The following is a checklist of sequential steps to follow to prepare these input parameters. .

1. Define the VA operational clusters units (see Part A, section 3.1).

2. Decide on the level of result disaggregation (see Part A, section 3.3)
3. Decide on the number of years to be aggregated for trend analysis (see Part, section A 5.3.5)
4. Identify the best source of national estimates for the top 20 proportionate causes of deaths (e.g. from a local source or IHME GBD Compare) (See Annexes C.2 and C.3).
5. Identify from these estimates, the top 20.
6. Check if the identified causes are included in the VA target cause list; if not continue down the list until 20 target causes are identified.
7. Note the range of CSMFs between the 1st and 20th cause.
8. Determine the best estimate of national annual CDR for the current year and get the same disaggregated by sex (see Part A, section 5.3.4 and Annex C.1).
9. Get national estimates for the proportion of total deaths with a MCCD or occurring within a health facility depending on your decision (see Part A, section 2.2 and 5.3.6).
10. Get an estimate from the pilot for the proportion of deaths for which a VA should be conducted, but is not. This is the under-notification and non-response rate (see Part A, section **Error! Reference source not found.** and 5.3.7).
11. Prepare a sampling frame database of the operational clusters as explained above (Part B, section 2).
12. As an optional step you could at this stage map the clusters using a GIS to identify eligible clusters based on population density thresholds for the final sampling frame by excluding outlier clusters at the extreme ranges of the population density (e.g. national parks with extremely low population density or refugee camps with extremely high population density). This will result in the final sampling frame with all eligible clusters (see Part A, section 4.2).
13. Calculate the mean cluster population for the remaining clusters in the sampling frame.
14. If possible, estimate the coefficient of variation of the true proportions between clusters, k (see Annex B.2).
15. If possible, estimate the Maximum possible Inflation in sample Size. MIS (see Annex B.4).
16. Enter required parameters into the CRVS VA Sample Size Calculator Tool to determine the number of clusters to be sampled for various scenarios of interest.

3.2 Using the CRVS VA Sample Size Tool

Once all preparatory steps are completed you can start using the CRVS VA Sample Size Tool to calculate the sample size required. To do so, follow the steps described here:

1. Open the CRVS VA Sample Size Calculator Tool in MS Excel version 2010 or later:
2. Enter the name of your country and the year.
3. Define whether you want to use the tool in Mode 1 or 2 and enter the required parameters.
    a. **Mode 1**: Calculate the required sample size for a pre-determined maximum acceptable uncertainty range for detecting significant CSMF changes over time (Part A, section 5.3.2): The uncertainty range is the range within which changes in CSMFs over time will not be detectable with sufficient statistical power. Note: the range is given in percentage change, which is not to be mistaken with percentage point change. We have indexed the Calculator to allow you to specify the acceptable uncertainty range for detecting significant CSMF changes for the smallest CSMF of interest (causes in the range of 1% of total deaths) as for this cause the range will be the widest. Uncertainty ranges at all higher CSMFs will always be narrower. Thus, the value to be entered is the uncertainty range (percentage change) of the 1% CSMF that is acceptable. For example, setting the acceptable uncertainty at 50% means that the tool will calculate a sample size capable to detect a change in CSMF over time of 50% or more, meaning for the 1% CSMF an increase to 1.5% or a decrease to 0.5% in the next time period.

   b. **Mode 2**: Calculate the uncertainty range for detecting significant CSMF changes over time for a given number of clusters (Part A, section 5.3.1): For this, you will need to enter the number of clusters you are expecting to sample.
4. Decide if you want to calculate your sample size in order to be able to disaggregate your results by male and female. Choose "Yes" or "No".
5. Select whether or not you know the population size of each eligible clusters. If you already prepared your final sampling frame (Part B, section 3.1), then this answer is yes.
6. If you know the population size of each eligible cluster, enter their names, population size and crude death rate (best to copy from your sampling frame). If you do not know the cluster specific crude death rate, use the best proxy, which might be the crude death rate from any administration level above the cluster level or simply the national crude death rate.
7. Enter your input parameters (Part A, section 5.3):
   a. Power (set to default 0.8 = 80%).
   b. Significance level (set to default 0.05 = 5%).
   c. Mean population per cluster (only if you do not know the population size of each eligible cluster).
   d. Mean crude death rate per 1,000 (only if you do not know the CDR of each eligible cluster).
   e. Estimation for the coefficient of variation in the true proportions between clusters, k (set to default 0.25).
   f. Estimation for the Maximum possible Inflation in sample Size, MIS (default 1.5 if population size of each eligible cluster is not known, otherwise 1).
   g. Number of years to aggregate for trend analysis.
   h. Adjustment for proportion of deaths having MCCD or occurring with a health facility.
   i. Adjustment for under-notification and non-response (%).
8. In the results section you will see the following output parameters (Part A, section 5.4):
   a. Number of clusters required.
   b. Estimated total population in the sample.
   c. Estimated number of deaths in the sample per year.
   d. Estimated number of VAs needed per year.
   e. Table of estimated uncertainty ranges for pre-set CSMFs levels (25%, 20%, 15%, 12.5%, 10%, 7.5%, 5%, 3%, 2%, and 1%).

      Note: CSMF levels are pre-set and you will need to remember what CSMFs levels are relevant for your context and only consider those.

      Note: You will need to make a decision about the number of clusters to be sampled based on the acceptable uncertainty range for detecting significant CSMF changes of ONE specific CSMF level (cause of death). For all other CSMF levels (causes of death) the uncertainty range will be given by the calculated number of sampled clusters.

9. You can now play with the uncertainty range or the number of clusters in a "what if" manner, depending on what you selected in step 3. This will allow you to observe how your input affects the number of clusters required and the uncertainty ranges of the CSMF levels.
10.

Note: In case you are also interested in mortality rates, the uncertainty ranges for detecting significant CSMF changes over time is very similar to the uncertainty range in rates assuming all other parameters being constant.

4. Selecting the sample clusters

Having determined the sample size in terms of the number of clusters needed, the sampling strategy now needs to specify how to select the sample clusters to be included in the CRVS VA system from the sampling frame. Recall the cluster unit is the defined catchment area of the VA interviewer or interviewer team. Within a cluster, VAs would need to be done for all deaths eligible for VA within the resident population of the cluster unit. The final sampling frame from which outlier clusters have been excluded (Part B, section 2) will be required. There are various types of sampling strategies. Here we will describe the procedures for the in Part A recommended "stratified single-stage proportional to size cluster sampling" (Part A, section 6.1).

4.1 Stratification

Depending on the circumstances in your country, it might be desirable to make sure that clusters are representative for the number of deaths occurring within each major Administrative Level 1 (e.g. Level 1, Region, State, and Province) and within urban and rural areas.

To do the above suggested stratification, follow the steps listed here:

1. Group the clusters in the sampling frame by administrative level 1 (region) and by urban/rural areas to create so-called 'strata'. For instance, urban clusters from region 1 fall into one stratum (stratum 1), rural clusters from region 1 into another (stratum 2) and so on.
2. Distribute the total number of required clusters across all the strata proportional to the population in each stratum. For instance, if all people from urban clusters in region 1 (stratum 1) make up 5% of the total population in the country, 0.05 times the total number of required clusters would be assigned to this stratum.
3. Select the required number of clusters within each stratum based on the below explained Probability Proportional to Size sampling processes.

4.2 Probability Proportional to Size sampling

To implement Probability Proportional to Size (PPS) sampling, do the following (Bierrenbach 2008):

1. List all clusters within a stratum with their name, unique ID and the number of estimated population in the cluster
2. In a new column, calculate the cumulative sum of the cluster population (the total population in the stratum should be the last figure in this column)
3. Obtain the number of clusters to be sampled (d) in each stratum (from the Sample Size Calculator Tool).
4. Divide the total population in the stratum by the number of required clusters (d) in the stratum, leading to the sampling interval, SI
5. Choose a random number between 1 and the sampling interval to get the random start, RS
6. Calculate the following series: RS; RS+SI; RS+2*SI; … RS+(d-1)*SI
7. The clusters selected are those for which the cumulative sum contains one of the serial numbers calculated in 6).

	A	B	C	D	E
1	**Cluster name**	**Population size**	**Cumulative sum**	**Selected**	
2	Cluster 1	11,594	11,594	0	
3	Cluster 2	10,325	21,919	1	
4	Cluster 3	7,457	29,376	0	
5	Cluster 4	5,271	34,647	0	
6	Cluster 5	6,370	41,017	0	
7	Cluster 6	9,697	50,714	0	
8	Cluster 7	16,079	66,794	1	
9	Cluster 8	9,009	75,803	0	
10	Cluster 9	17,218	93,021	0	
11	Cluster 10	6,468	99,490	0	
12	Cluster 11	12,630	112,120	0	
13	Cluster 12	6,783	118,903	0	
14	Cluster 13	13,119	132,022	1	
15	Cluster 14	13,847	145,869	0	
16	Cluster 15	2,904	148,772	0	
17	Cluster 16	2,094	150,867	0	
18	Cluster 17	8,109	158,976	0	
19					
20					
21	**Total population size**		158,976		
22	**Required clusters in the stratum**		3		
23	**Sampling interval**		52,992		
24	**Random number**		13,577		
25	**Series**		13,577; 66,569; 119,561		
26					

ANNEXES

Annex A. Statistical basis of the CRVS VA sample size calculations

A.1. Assumptions

Sample size estimates are driven by the sampling strategy. Here we assume a sampling strategy which uses a **single-stage random cluster sampling design** (see Part A, Section 6.1). The cluster unit is assumed to be the defined catchment area of the VA interviewer or interviewer team as suggested in Part A, Section 3.1. Further, we expect at a minimum, that results must be **representative for the national level** and **disaggregated by sex**. We also set the number of years to aggregate for a trend analysis to three. The clusters selected for the first three years, will be the same in the second three years. This means for the sample size calculations you need to calculate the number of clusters needed to detect a significant change in one CSMF between the first and the second three year periods. All input parameters are assumed constant over time.

A.2. Power and Significance level

In a first step, you need to define to what extent you can accept statistical errors in such a trend analysis. There are two types of statistical errors. The first would be that you conclude a change in one CSMF between the first and the second three year periods, although there is no such change in reality. By defining the **significance level (referred to as α),** you decide on an acceptable probability of making this kind of error. This value is also known as the p-value and is often set to 0.05 (5%).

The second error would be that you conclude no change in one CSMF between the first and the second three years, although there is such a change in reality. An acceptable probability of doing this error is generally seen to be 0.2 (20%) **(referred to as β)**. The **power** of a statistical test is consequently the probability of not making this error and therefore is often set to 0.8 (80%).

This means you assume that the sample size is designed to give 80% power of obtaining a significant difference ($p<0.05$) for a given change in CSMF.

A.3. Individual vs. Cluster design

In an **individual design**, you would randomly select deaths (individuals) from an area, in this case the country, to analyze the CSMFs. To do so you would need to know how many deaths (individuals) you would need to analyze in order to detect a significant change in one CSMF between the first and the second three years. This is logistically infeasible and therefore you will need to select a number of clusters within which you analyze all deaths (**cluster design**). Thus, instead of calculating the number of deaths (individuals) to be sampled, you need to calculate the number of clusters to be sampled from the sampling frame.

A.4. Unmatched vs. Matched design

In an **unmatched design** two groups of observations are compared, which originate from a completely different set of individuals or clusters. This would for example be the case if you randomly selected a first group of clusters for the first three years and another group of clusters for the second three years and then compared the CSMFs of the two different groups of clusters.

In a **matched design** two groups of observations are compared, which are paired, meaning they originate from the same or similar set of individuals or clusters. Matching should lead to greater comparability between the two groups of observations (Hayes and Bennett 1999). This is the case in your scenario where the clusters selected for the first three years will be the same in the second three years. Thus, observations from the first three years are matched with observations of the second three years. This assumes that the clusters in the second three years are still representative.

A.5. Design related parameters

The fact that the design is clustered and these clusters are paired and of unequal size requires you to take into account some additional parameters when calculating the sample size.

A.5.1.Coefficient of variation between clusters

In an unmatched design, you would need to take into account that there is some variation in CSMFs between clusters. This variation is given by k, the coefficient of variation of true proportions (or rates) between clusters at each time point (in the first three years or second three years) (Hayes and Bennett 1999). k is defined as the standard deviation of true proportions (or rates) divided by the mean proportion (rate) (Hayes and Bennett 1999). If there is no variation in one CSMF between clusters then k would be 0 (Hayes and Bennett 1999). As a rough guideline, Hayes and Bennett state that experience suggests that k is often $\leq$0.25, and seldom exceeds 0.5 for most health outcomes (Hayes and Bennett 1999). Options for how to estimate k are given in Annex B.3.

In a matched design, k is replaced by km. k_m is the coefficient of variation of true proportions (or rates) between clusters **within the matched pairs in absence of** anything which could change the mortality and/or the CSMFs (Hayes and Bennett 1999). This means that k_m is the coefficient of variation in one CSMF between the first three years and the second three years within a cluster in absence of anything which could change the mortality and/or the CSMF. This might for example only include in- and out-migration and ultimately lead to the question how similar the populations are in the second three years compared to the population in the first three years. It is impossible to obtain empirical estimates of k_m in practice and thus we propose the **conservative assumption of k (= 0.25) being the upper limit for k_m**[6] as recommended by Hayes and Bennett (Hayes and Bennett 1999). Options to estimate k_m based on non-empirical estimates are given in Annex B.3.

A.5.2.Intra-cluster correlation coefficient

In a cluster design there are two types of variances – the variance of observation within the same cluster and the variance of true cluster means (Kerry and Bland 1998). One way of summarizing the relationship between these two components is the Intra-cluster Correlation Coefficient (ICC) (Kerry and Bland 1998). The ICC is defined as the ratio of the between-cluster variance to the total variance (both between and within clusters) (Pagel, Prost et al. 2011). Thus, it quantifies how much more similar outcomes are for individuals within clusters than for those in different clusters (Kerry and Bland 1998, Killip, Mahfoud et al. 2004, Pagel, Prost et al. 2011). It has a value between 0 and 1 (Pagel, Prost et al. 2011). An ICC of 0 means that observations within clusters are no more similar to each other than observations from different clusters (there is no between-cluster variability) (Pagel, Prost et al. 2011). In contrast, an ICC of 1 indicates that observations within the same cluster all have identical outcomes (there is no within-cluster variability) (Pagel, Prost et al. 2011). For binary outcomes, the relationship between the ICC and k has been defined as follows:

$$ICC = k^2 \left(\frac{\pi}{1-\pi}\right)$$

whereas π in your case would be the probability of dying of one specific cause (the CSMF for one specific cause) (Preisser, Reboussin et al. 2007, Pagel, Prost et al. 2011).

A.5.3.Design effect

According to Eldridge et al. the design effect, DE, represents the amount by which the sample size required for a non-cluster design (individual design) needs to be multiplied to obtain the size required for a more complex design such as a cluster design (Eldridge, Ashby et al. 2006). Pragmatically, the DE is the number of deaths (individuals) needed in a cluster design (n2) divided by the numbers of deaths (individuals) needed in a non-cluster design (n1). The DE can also be formulated using the ICC, whereas the commonly used DE estimate for equal cluster sizes is given by:

[6] Alternatively, a less conservative assumption would be to use k=0.15.

$$DE = 1 + (m - 1)ICC$$

with m being the average size of the cluster (number of deaths for CSMFs) (Hayes, Alexander et al. 2000, Eldridge, Ashby et al. 2006, Pagel, Prost et al. 2011).

A.5.4.Coefficient of variation in cluster size

Variation in cluster size needs to be taken into account in addition to the variation of outcomes between clusters (k or km) (Eldridge, Ashby et al. 2006). To do so you need to know the coefficient of variation of cluster size, cv, which is defined as the ratio of the standard deviation of cluster sizes to the mean cluster size (Eldridge, Ashby et al. 2006). Together with the ICC, you can then calculate the Maximum possible Inflation in sample Size (MIS) required when cluster sizes are variable rather than equal.

$$MIS = \frac{1 + [(1 + cv^2)m - 1]ICC}{1 + (m - 1)ICC}$$

An alternative to the MIS would be to use the harmonic mean of the cluster size instead of the arithmetic mean m and y in the formula in A.6 and A.7 (Hayes and Moulton 2009). Yet, for this you need to know at least the population size of each eligible cluster. If you do, select the corresponding option in the Sample Size Calculator Tool and the tool will consider the harmonic mean. In this case the MIS will be automatically set to 1, otherwise the MIS is set to 1.5. An MIS of 1.5 was chosen because this revealed a similar number of clusters required as if the harmonic mean was used all other parameters being constant.

A.6. Formula for sample size based on proportions

CSMFs represent proportions of one specific cause of death compared to all deaths. According to Hayes and Bennett (Hayes and Bennett 1999) the sample size formula for proportion (CSMFs) in matched designs is as follows:

$$c = 2 + (z_{\alpha/2} + z_{\beta})^2 \frac{\pi_0(1 - \pi_0)/m + \pi_1(1 - \pi_1)/m + k_m^2(\pi_0^2 + \pi_1^2)}{(\pi_0 - \pi_1)^2}$$

c is the number of clusters required. $z_{\alpha/2}$ and z_{β} are standard normal distribution values corresponding to $\alpha/2$ and β respectively (see Annex A.2) (Hayes and Bennett 1999). π_0 is the true mean population proportion for a specific cause in the first three years and π_1 is the true mean population proportion of a specific cause in the second three years. m is the number of individuals, in your case deaths, sampled in each cluster (average size of the cluster using the arithmetic mean). As discussed above k_m will need to be replaced by k (Annex A.5.1).

A.7. Formula for sample size based on rates

Based on Hayes and Bennett (Hayes and Bennett 1999) the sample size formula for rates, in your case mortality rates, in matched designs is as follows:

$$c = 2 + (z_{\alpha/2} + z_{\beta})^2 \frac{(\lambda_0 + \lambda_1)/y + k_m^2(\lambda_0^2 + \lambda_1^2)}{(\lambda_0 - \lambda_1)^2}$$

λ_0 is the true mean population rate for a specific cause in the first three years and λ_1 is the true mean population rate of a specific cause in the second three years. y is the number of person-years of follow-up in each cluster (average size of the cluster using the arithmetic mean).

A.8. Cluster size

CSMFs represent the proportion of one cause of death in all deaths occurring. Thus, the **denominator is all deaths**. For mortality rates, you look at the number of deaths of one cause occurring in a population, which is often expressed as x deaths per 100,000 person-years. In this case the **denominator is person-years**. This differentiation is also important for the above formulas.

The cluster size m in the formula for proportions is the number of deaths in each cluster for the period under investigation. This means to calculate the cluster size m you have to multiply the average number of deaths per year and cluster by the number of years, which you will aggregate for a trend analysis.

In contrast, the cluster size y in the formula for rates is the number of person-years of follow up in each cluster for the period under investigation. To calculate the cluster size y you would therefore need to multiply the average number of individuals per cluster (number of person-years of follow up per year) by the number of years, which you will aggregate for a trend analysis.

A.9. Uncertainty range

Both the formulas for proportions and rates in A.6 and A.7 require you to know the true proportion or rate of a specific cause in the second three years. For this, you would at least need to know the expected change from the first three years to the second three years. Because this is not possible, the Sample Size Calculator Tool will require you to put in an acceptable uncertainty range. The uncertainty range is the range within which changes in CSMFs over time will not be detectable with sufficient statistical power. The range is given in percentage change, which is not to be mistaken with percentage point change. The Sample Size Calculator Tool also allows you to play with this uncertainty range, meaning the minimal percentage change you would like to be able to detect, in a "what if" manner and observe how it affects the number of clusters needed.

A.10. Further Adjustments

Once you calculated the c (number of required clusters) in the above formulas (A.6 and A.7), there are still a couple of adjustments needed to account for additional factors influencing the sample size.

A.10.1. Disaggregation by male and female

In order to have disaggregated results for male and female, you will need roughly to double the calculated number of clusters required. In theory, this estimation only holds true if the CDRs for female and male are somewhat similar. If you wish to know if this is the case in your situation, you would need to calculate the required number of clusters for the sub-population of the sex with the lower CDR (often female). An example of how this is ought to be done is attached in Annex B.7.

A.10.2. Proportion of deaths having MCCD

If the CRVS VA strategic design decision is to conduct VA only on deaths which had not had a MCCD or do not occur within a health facility, the number of deaths that can be analysed in each cluster will be smaller. Thus, the cluster size m in the formula for proportions or y in the formula for rates will be smaller. This will increase the numbers of clusters needed. The factor by which the cluster size has to be adjusted is given by

$$Factor_{MCCD} = (1 - p_{MCCD})$$

where p_{MCCD} is the proportion of deaths for which a MCCD is available or the proportion of deaths which occur in a health facility.

Again, this calculation only holds true if the proportion of deaths with an MCCD or occurring in a health facility is similar across the country. If this is not the case (e.g. if in urban areas the proportion of deaths with an MCCD is much higher than in rural areas), then you will need to calculate the number of clusters required in urban and rural areas separately. An example of how to do this is given in Annex B.8.

A.10.3. Under-notification and non-response rate

As introduced in Part A, section 5.3.7, the proportions of deaths for which VA should be conducted, but is not, is the so-called under-notification and non-response rate. Again, the factor by which the cluster size has to be adjusted is given by

$$Factor_{UN\&NR} = (1 - p_{UN\&NR})$$

where $p_{UN\&NR}$ is the under-notification and non-response rate.

Annex B. Worked example implementing the Guidance and Tool in one country

For illustrative purpose of this guidance document, Tanzania was selected to provide a worked example of how to calculate and draw the CRVS VA sample in practice. For this, we will follow the steps outlined in Part B.

B.1. Preparing to calculate the Cluster Sample Size

B.1.1. Operational cluster definition

VA implementation will take place in Tanzania Mainland only (excluding Zanzibar). The administration level 1 in Tanzania Mainland is the region, level 2 the district, level 3 the division, and level 4 the ward. The catchment area of a VA interviewer team in Tanzania, which consists of two VA interviewers, is the ward (Administration area level 4). Thus, we decide **the cluster unit to be defined as the ward**. According to the census 2012, there are 3,312 wards in Tanzania Mainland with an expected total population of 50,366,198 in 2017 (extrapolated based on inter-censal administration level 2 growth rates from the 2012 census; see B.2) (National Bureau of Statistics and Ministry of Finance 2013). The wards vary widely in area from 0.1 to 11,503 km^2. Also, ward population sizes vary considerably from 753 to 148,017 people. Given an annual national CDR of 6.351/1,000 for 2017, there are on average 97 deaths per ward per year. (You do not need all this information at this point, but it is provided here to give you a brief overview of the situation if you are not familiar with Tanzania)

B.1.2. Disaggregation level of results

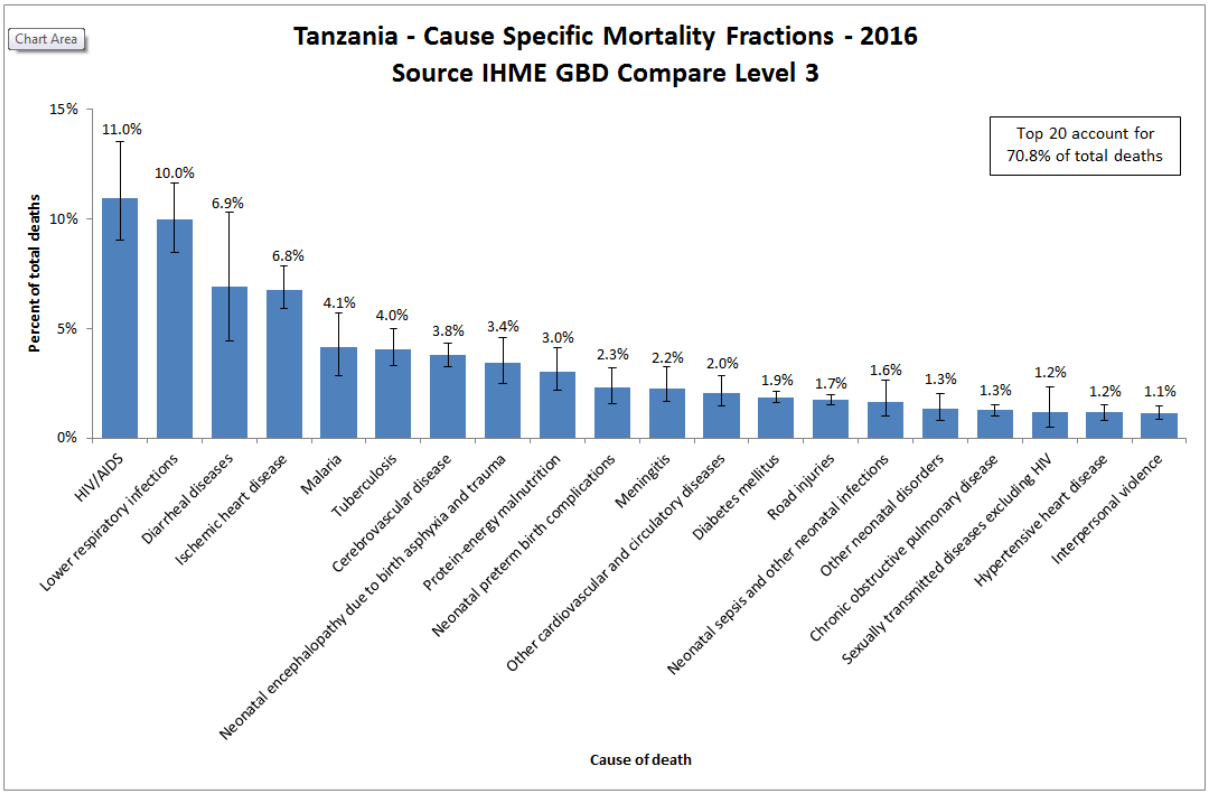
For the disaggregation level of results, it is decided that the results ought to be **representative for males and females at national level**.

B.1.3. Number of years to aggregate for trend

In Tanzania a step-wise up-scaling is planned and full scale sample VA operations are expected to be reached only after a couple of years. We will **allow for a three year follow-up period for national trends**. This means that data will be gathered over three years and then compiled to one data set. In long-term, it is aimed to compare a three year period with the subsequent three years.

B.1.4. National level estimates

To get the **estimated CSMFs** for Tanzania download the CSMFs data from https://vizhub.healthdata.org/gbd-compare/ and prepare a frequency distribution as described in Annex C.2. As Tanzania uses the WHO 2016 VA cause list we will drop “Alzheimer disease and other dementias”, which is ranked as the 18th cause. Instead we add the 21st cause, which is “Interpersonal violence”.



We note that the CSMFs range between 11% and 1.1%.

To get the **annual national overall CDR**, we conduct an online search, which reveals:

CDR [per 1'000]	Year	Source	Link
6.351	2017	World Population Review[7]	http://worldpopulationreview.com/countries/tanzania-population/crude-death-rate/
6.351	2015 - 2020	UN/population division[8]	https://esa.un.org/unpd/wpp/Download/Standard/Mortality/
6.49	2017	Knoema	https://knoema.com/atlas/United-Republic-of-Tanzania/topics/Demographics/Mortality/Crude-death-rate
6.68	2015	Index Mundi	http://www.indexmundi.com/facts/tanzania/indicator/SP.DYN.CDRT.IN
6.68	2015	Country Economy	https://countryeconomy.com/demography/mortality/tanzania
7.015	2015	World Bank[9]	https://data.worldbank.org/indicator/SP.DYN.CDRT.IN?locations=TZ&view=chart

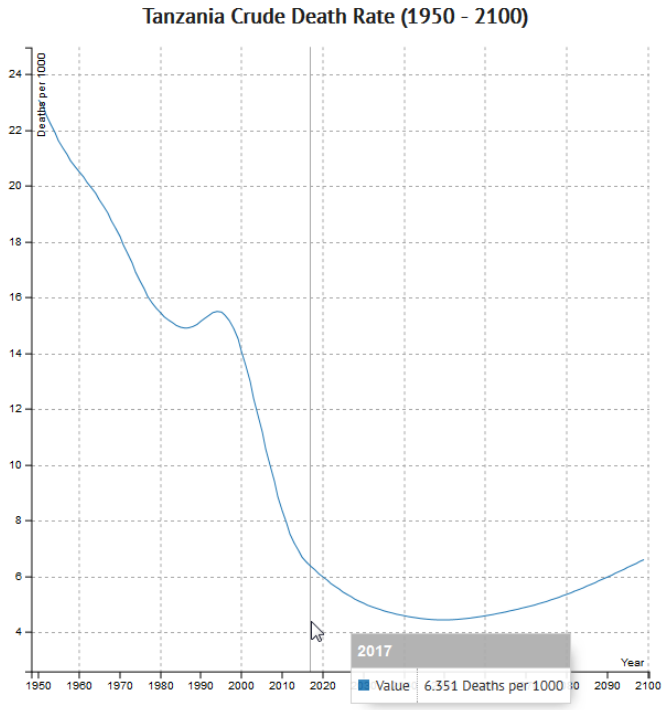
[7] http://worldpopulationreview.com/countries/xxx-population/crude-death-rate/ (xxx has to be replaced by the country's name).

[8] UN: https://esa.un.org/unpd/wpp/Download/Standard/Mortality/

[9] https://data.worldbank.org/indicator/SP.DYN.CDRT.IN?locations=XX (xxx has to be replaced by the country's name)

7.6	2017	CIA	https://www.cia.gov/library/publications/the-world-factbook/geos/tz.html
7.79	2015	Open Data For Africa	http://tanzania.opendataforafrica.org/UNWPP2012R/world-population-prospects-the-2012-revision-updated-13-june-2013
7.8	2013	WHO	http://apps.who.int/iris/bitstream/10665/170250/1/9789240694439_eng.pdf
8.8	2012	UNICEF	https://www.unicef.org/infobycountry/tanzania_statistics.html
9.4	2012	NBS/Census	http://www.nbs.go.tz/nbs/takwimu/census2012/Mortality_and_Health_Monograph.pdf

Here is one example from World Population Review: Modeled trend in crude death rate for Tanzania, 1950 to 2100



Given that we expect a further decrease in the CDR in the coming years, we select the lowest CDR of 6.351. For **CDRs disaggregated by sex,** data from the 2012 census for Mainland Tanzania is available. Assuming that the ratio of the overall CDR to the male and female CDR remained the same, we can calculate the expected CDRs for male and female in 2017 as follows:

Year	Overall CDR	Male CDR	Female CDR
2012	9.4	10.1	8.6
2017	6.351 (from above)	=10.1/9.4*6.351=7.43	=8.6/9.4*6.351=5.81

Here we assume that in Tanzania the CRVS VA strategic design decision is to conduct VA only on deaths, which had not had a MCCD. Thus, we need an **estimate of the percentage of total deaths with a MCCD**. We know

from DHIS2 that in 2016 34,189 deaths had a MCCD. Additionally, the population projection from the National Bureau of Statistics estimated a total of 48,676,699 people on Tanzania Mainland in 2016. With this and an expected CDR of 6.502 (Source: World Population Review) for 2016, we would expect 316,495 deaths in 2016. Comparing the expected number of deaths for 2016 and the deaths with a MCCD in 2016, reveals that only **11%** of all deaths have a MCCD. Thus, we will use this estimate for further calculations later on.

Regarding the under-notification and non-response rate, we here assume to have a notification achievement level of 90% and a response rate close to 100%. This will lead to an **under-notification and non-response rate of 0.1 (10%)**, which is rather optimistic. If possible, this rate should be carefully analyzed using results from the areas to set this parameter correctly.

B.2. The sampling frame

To prepare the sampling frame of the format given in Part B, section 2, for Tanzania Mainland, we use shape files and additional data from the census 2012.

The **estimated population size for 2017**, (EstPop2017), is calculated with the following formula:

$$EstPop2017 = Pop2012 * e^{\left(\frac{r}{100} * t\right)}$$

whereas Pop2012 is the population size in 2012, r the inter-censal growth rate per annum of a corresponding district and t the number of years between 2012 and 2017 (in this case t=5).

The **area** we calculate using the QGIS function $area/1,000,000.

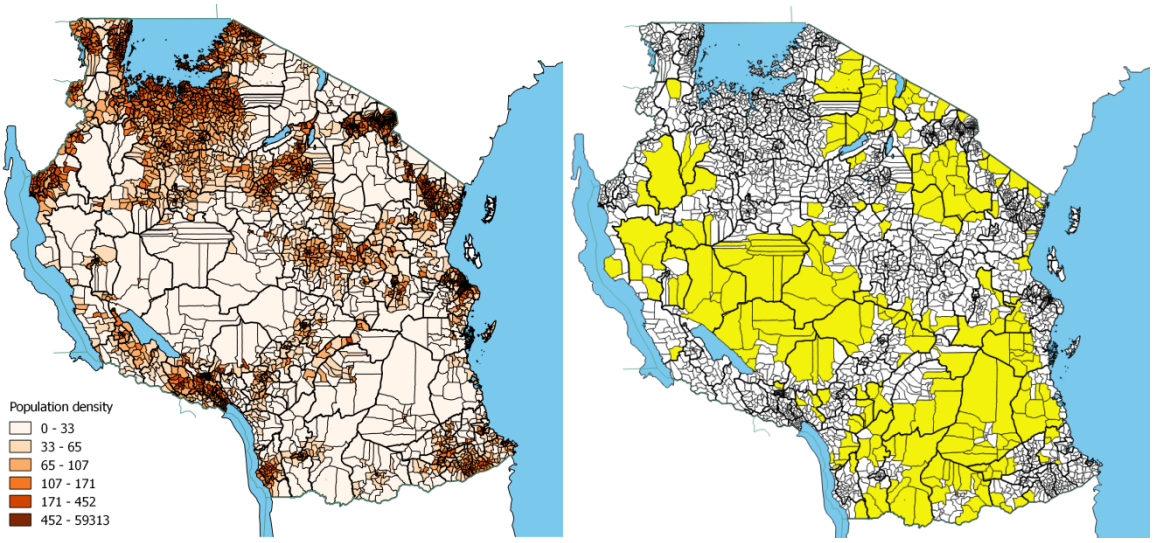
In the 2012 census the CDR was available down to regional level. Similar to calculating the sex specific CDR, we calculate the **estimated regional CDR for 2017** based on the ratio of the regional CDR 2012 to the national CDR 2012 and multiplying this ratio by the 2017 CDR.

	A	B	C	D
1	**CDR calculations**	***2012***		***2017***
2		***CDR***	***Ratio to national*** (=B/$B$3)	***CDR*** (=C*$D$3)
3	Mainland average	9.4	1	6.351
4				
9	Singida	7.8	0.82978723	5.27
10	Mwanza	8.4	0.89361702	5.68
11	Dodoma	9.0	0.95744681	6.08
12	DaresSalaamRegion	9.1	0.96808511	6.15

The **estimated expected number of deaths in 2017** is then the estimated population in 2017 times the estimated regional CDR for 2017 divided by 1,000.

As a next step we visualize the 3,321 wards of Tanzania Mainland using QGIS to look at population density (left map) and in particular to find a reasonable population density threshold until which VA implementation is still feasible.

For example using 15 people/km^2 as a cut off would drop 236 of the 3,312 wards, whereas within those there are 80 of the 83 biggest wards in terms of area (right map). This makes sense regarding feasibility as in wards with a big area VA implementation will be challenging.



Dropping these 236 leads to the fact that we drop 4.4% of total population. The following table summarizes the characteristics of the sampling frame **with and without the wards with a population density lower than 15/km²**.

Relevant characteristics	**All wards in Tanzania Mainland**	**Excluding wards with a population density <15/km2**
Number of wards	3,312	3,076
Area [km²]	mean=270km²; median=112 km²; min=0.1 km²; max=11,503 km²	mean=160 km²; median=103 km²; min=0.1 km²; max=1,728 km²
Total estimated population in 2017	50,366,198	48,140,318 (95.6%)
Estimated population per ward in 2017	mean=15,207; median=11,825; min=753; max=148,017	mean=15,650; median=12,185; min=1,137; max=148,018
Estimated expected death per ward in 2017	95.96749 (arithmetic mean)	98.70358 (arithmetic mean) 64.97327 (harmonic mean)
Average number of villages per ward	4.6	4.7

We decide that the sampling frame without the low population density wards, will be our final sampling frame. In B.5 we will need the **mean population per ward, which is 15,650** and the **mean number of deaths per ward in 2017, which is 98.7 (based on the arithmetic mean) or 65.0 (based on the harmonic mean)**.

B.3. Estimating k

As elaborated in A.5.1 k is the coefficient of variation in the outcome measure between clusters at one time point. Here, we will give you some options how to estimate the k based on existing data.

B.3.1. Options to estimate k

Option 1: In Tanzania we know the CDR at regional level, which is three administration levels above our cluster level (the ward). However, we will here calculate the coefficient of variation in CDRs between regions, k_{REGION}. Assuming that variation between clusters increases the smaller the clusters are and that cause specific mortalities vary at least as much as the overall CDR, this estimated k_{REGION} will give us some idea of the $k_{CLUSTER's}$ minimal value.

According to Hayes and Moulton k is defined as follows (Hayes and Moulton 2009):

$$k = \frac{\sigma_B}{\pi_0} \; or \; k = \frac{\sigma_B}{\lambda_0}$$

Whereas the σ_B is the between-cluster standard deviation and the π_0 or λ_0 is the mean population proportion and rate, respectively. In our case the k is k_{REGION}, σ_B is the between-region standard deviation, and λ_0 is the mean population rate.

	A	B	C	D	E
1	**Region**	**Estimated regional CDR (2017) [/1,000]**	**Estimated regional CDR (2017) (=B/1,000)**	**Estimated Population (2017)**	**Estimated deaths (2017) (=C*D)**
2	Arusha	3.728298	0.00372830	1'766'560	6'586
3	DaresSalaam	6.282873	0.006282873	5'786'614	36'357
4	Dodoma	6.21383	0.00621383	2'248'377	13'971
5	Geita	5.247234	0.005247234	1'960'259	10'286
6	Iringa	8.63032	0.00863032	894'158	7'717
7	Kagera	7.870851	0.007870851	2'923'821	23'013
8	Katavi	7.387553	0.007387553	444'058	3'281
9	Kigoma	6.628085	0.006628085	2'367'907	15'695
10	Kilimanjaro	6.420958	0.006420958	1'762'457	11'317
11	Lindi	7.04234	0.00704234	769'764	5'421
12	Manyara	4.763937	0.004763937	1'511'831	7'202
13	Mara	6.559043	0.006559043	1'940'001	12'725
14	Mbeya	7.387553	0.007387553	3'011'260	22'246
15	Morogoro	6.351915	0.006351915	2'372'681	15'071
16	Mtwara	7.318511	0.007318511	1'317'875	9'645
17	Mwanza	5.799574	0.005799574	3'235'643	18'765
18	Njombe	9.320745	0.009320745	657'136	6'125
19	Pwani	8.423191	0.008423191	1'142'996	9'628
20	Rukwa	6.973298	0.006973298	1'196'437	8'343
21	Ruvuma	6.835212	0.006835212	1'264'163	8'641
22	Shinyanga	6.697127	0.006697127	1'708'973	11'445
23	Simiyu	5.109149	0.005109149	1'661'813	8'490
24	Singida	5.385319	0.005385319	1'469'017	7'911
25	Tabora	6.559043	0.006559043	2'466'906	16'181
26	Tanga	6.282873	0.006282873	2'259'613	14'197
27	**SUM**	**n/a**	**n/a**	**48'140'318**	**310'257**
28					
29	**Mean population rate (=E27/D27)**				**0.006444851**
30	**Between-region sandard deviation (=STDEV.S(C2:C26))**				**0.001239168**
31	**k_{REGION} (=E30/E29)**				**0.192272562**
32					

You will need to fill cells highlighted in yellow based on your input data.

If you had CDRs for all of your clusters, you could do the same calculations as done here for regions and get your true $k_{CLUSTER}$ fro the overall CDR.

Option 2: If the true cluster proportions or rates are approximately normally distributed, 95% will lie within two standard deviations of the population mean (Hayes and Bennett 1999). For a given k this would imply that the rates (or the proportions) in the clusters would vary roughly between (Hayes and Bennett 1999):

$$\lambda_0 * (1 \pm 2 * k_{CLUSTER})$$

where λ_0 is the mean population rate (for proportions λ_0 would need to be replaced by π_0). This means we need to choose $k_{CLUSTER}$ big enough so that the above term captures at the minimum the ranges we see in reality. The IHME data downloaded under B.1.4 give us a rough idea what the "Lower Bound" and "Upper Bound" could be. Thus, we now use these bounds to estimate the required k. Ultimately, the following would need to be true:

$$\lambda_0 * (1 - 2 * k_{CLUSTER}) < Lower\ Bound_{IHME}$$
$$\lambda_0 * (1 + 2 * k_{CLUSTER}) > Upper\ Bound_{IHME}$$

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Rank	Rate [/100,000]	Lower Bound [/100,000]	Upper Bound [/100,000]	Rate [λ] (=B/100,000)	Lower Bound (=C/100,000)	Upper Bound (=D/100,000)	λ*(1-2*k) (=E*(1-2*$B$23))	λ*(1+2*k) (=E*(1+2*$B$23))	Check (=IF(OR(I>F;J<G); "false";"true"))		Rank	CSMF [π]	Lower Bound	Upper Bound	π*(1-2*k) (=M*(1-2*$B$23))	π*(1+2*k) (=M*(1+2*$B$23))	Check (=IF(OR(P>N;Q<O); "false";"true"))
2	1st	77.973321	66.317656	95.519624	0.000779733	0.0006632	0.0009552	0.0003899	0.0011696	true		1st	0.1095028	0.0905122	0.1356083	0.0547514	0.1642542	true
3	2nd	71.234351	57.571189	85.812951	0.000712344	0.0005757	0.0008581	0.0003562	0.0010685	true		2nd	0.099818	0.084547	0.1163581	0.0499090	0.1497270	true
4	3rd	49.152711	31.085384	73.881714	0.000491527	0.0003109	0.0007388	0.0002458	0.0007373	false		3rd	0.0688956	0.0442627	0.1031231	0.0344478	0.1033434	true
5	4th	48.397489	40.432444	56.92151	0.000483975	0.0004043	0.0005692	0.0002420	0.0007260	true		4th	0.0678354	0.0589381	0.0786803	0.0339177	0.1017531	true
6	5th	29.423329	19.850889	41.475912	0.000294233	0.0001985	0.0004148	0.0001471	0.0004413	true		5th	0.0412303	0.0286015	0.0572418	0.0206151	0.0618454	true
7	6th	28.847724	23.252181	36.416292	0.000288477	0.0002325	0.0003642	0.0001442	0.0004327	true		6th	0.0404095	0.0332955	0.0497703	0.0202047	0.0606142	true
8	7th	26.868648	21.977022	31.834667	0.000268686	0.0002198	0.0003183	0.0001343	0.0004030	true		7th	0.0376573	0.0323326	0.0433047	0.0188287	0.0564860	true
9	8th	24.56038	17.819065	33.362356	0.000245604	0.0001782	0.0003336	0.0001228	0.0003684	true		8th	0.0344164	0.0250379	0.0460526	0.0172082	0.0516246	true
10	9th	21.520533	15.342621	29.799752	0.000215205	0.0001534	0.0002980	0.0001076	0.0003228	true		9th	0.0301378	0.0219273	0.0414057	0.0150689	0.0452067	true
11	10th	16.309314	11.074024	23.247437	0.000163093	0.0001107	0.0002325	0.0000815	0.0002446	true		10th	0.0228448	0.0156379	0.0318556	0.0114224	0.0342673	true
12	11th	15.904983	11.761962	23.083055	0.00015905	0.0001176	0.0002308	0.0000795	0.0002386	true		11th	0.0222941	0.0167225	0.0324057	0.0111470	0.0334411	true
13	12th	13.281421	11.099205	15.53581	0.000132814	0.0001110	0.0001554	0.0000664	0.0001992	true		12th	0.0186146	0.0162233	0.0210497	0.0093073	0.0279219	true
14	13th	12.360429	10.415107	14.597555	0.000123604	0.0001042	0.0001460	0.0000618	0.0001854	true		13th	0.0173209	0.0152583	0.0195843	0.0086605	0.0259814	true
15	14th	11.77452	7.0766299	18.785875	0.000117745	0.0000708	0.0001879	0.0000589	0.0001766	false		14th	0.0164967	0.0100817	0.026168	0.0082484	0.0247451	false
16	15th	9.0985503	7.2187274	11.130603	9.09855E-05	0.0000722	0.0001113	0.0000455	0.0001365	true		15th	0.012754	0.0101192	0.0153725	0.0063770	0.0191310	true
17	16th	8.6734345	6.9829465	10.839508	8.67343E-05	0.0000698	0.0001084	0.0000434	0.0001301	true		16th	0.0121646	0.0099154	0.0149194	0.0060823	0.0182469	true
18	17th	8.417476	3.2345175	16.623012	8.41748E-05	0.0000323	0.0001662	0.0000421	0.0001263	false		17th	0.0117972	0.0046832	0.0232743	0.0058986	0.0176958	false
19	18th	8.2503669	5.4699393	11.210251	8.25037E-05	0.0000547	0.0001121	0.0000413	0.0001238	true		18th	0.011559	0.0078189	0.0153072	0.0057795	0.0173385	true
20	19th	7.9931668	6.0140515	10.656946	7.99317E-05	0.0000601	0.0001066	0.0000400	0.0001199	true		19th	0.0111963	0.0086055	0.0144421	0.0055981	0.0167944	true
21	20th	7.2912374	5.924983	9.1455465	7.29124E-05	0.0000592	0.0000915	0.0000365	0.0001094	true		20th	0.0102182	0.0085758	0.0125941	0.0051091	0.0153273	true
22																		
23	k	0.25																
24																		

You will need to fill cells highlighted in yellow based on the downloaded IHME data. Columns J and R will notify you with “false” if the lower bound of the IHME data is lower than what is captured by your selected k or if the upper bound of the IHME data is bigger than what is captured by your selected k. You can now increase or decrease k in order to see what value for k captures most of the ranges given by the IHME data. Here we decide to select k = 0.25.

With “Conditional Formatting” in the “Home” tab of Excel, you can also highlight cells containing “false” and whether the lower, upper or both bounds lead to the “false” statement. To do so do the following:

Home > Conditional Formatting > New Rule > Format only cells that contain > select “Cell Value” & “greater than” (for Column H and P)/ “Cell Value” & “less than” (for Column I and Q)/“Specific Text” & ”containing” (for column J and R) > type “F2” (for column H), “G2” (for column I), “false” (for column J), “N2” (for column P), “O2” (for column Q), “false” (for column R) > click on “Format” > select the tab “Fill” > select the red color

Option 3: If you had the overall CDR or cause-specific mortality rates for a given number of clusters (here referred to as sample clusters) from a pilot or a similar intervention (e.g. SAVVY) you could use this data to get an idea of the k in all your clusters.

According to Hayes and Moulton (Hayes and Moulton 2009) the following holds true:

$$\sigma_B^2 = s^2 - \frac{r}{\bar{y}_H} \; and \; k = \frac{\sigma_B}{r}$$

where σ_B is the between-cluster standard deviation from all your clusters, s the standard deviation of the observed rates across your sample clusters, r is the overall rate of your sample clusters, and $\bar{y}_H$ the harmonic mean of the person-years in the sample cluster.

We did not have such data for Tanzania, but given this could be the case for other countries, we provide here an "artificial" example.

	A	B	C	D
1	**Sample clusters**	**Population**	**Annual deaths (=B*D)**	**CDR (=C/B)**
2	Ward 1	14'287	83	0.00580
3	Ward 2	7'618	71	0.00932
4	Ward 3	18'327	154	0.00842
5	Ward 4	15'690	109	0.00697
6	Ward 5	11'042	75	0.00684
7	Ward 6	19'090	128	0.00670
8	Ward 7	11'980	61	0.00511
9	Ward 8	4'545	24	0.00539
10	Ward 9	14'280	94	0.00656
11	Ward 10	11'583	73	0.00628
12	**Total**	**128'441**	**873**	**n/a**
13				
14	**Mean population rate (=C12/B12)**			0.006797555
15	**Standard deviation of rates across wards (=STDEV.S(D2:D11))**			0.001298702
16	**Harmonic mean of person-years per ward (=HARMEAN(B2:B11))**			10'841
17	**Variance (=D15^2-D14/D16)**			0.0000011
18	**k (=SQRT(D17)/D14)**			0.15
19				

You need to fill cells highlighted in yellow based on your data from the sample clusters. You will only need the population size (column B) and either the annual deaths (column C) or the CDR (column D). [10]

If you were to do the calculation for cause-specific mortality rates, you select one specific cause and then put in the cause-specific mortality rate instead of the CDR.

For proportions, meaning CSMFs, the above formula would be (Hayes and Moulton 2009):

$$\sigma_B^2 = s^2 - \frac{p(1-p)}{\bar{m}_H} \; and \; k = \frac{\sigma_B}{p}$$

where p is the overall proportion of your sample clusters (computed from all sample clusters combined) and $\bar{m}_H$ is the harmonic mean of the number of deaths in the sample clusters.

[10] Population sizes of Ward 1 to 10 in the example originate from 10 randomly picked wards in Tanzania and CDRs come from estimated regional CDRs in 2017 for Tanzania. Annual deaths were calculated correspondingly.

In conclusion, given the results from option 1 and 2 and the fact that k is often ≤0.25, **we perceive k=0.25 as appropriate, although potentially rather conservative**. Overall, if available data is insufficient to calculate k, it is better to overestimate k than underestimate it.

B.4. Estimating MIS

Here we provide an option on how to estimate the MIS. However, for sample size calculations we will use the harmonic mean of the cluster size instead of the arithmetic mean m and y in the formula in A.6 and A.7 (Annex A, Section A.5.4) (Hayes and Moulton 2009). We will base our calculations of the MIS on CSMFs. To calculate the MIS we need the ICC, the coefficient of variation in cluster size, cv, and the mean cluster size. The mean cluster size is the mean number of deaths in a cluster and known to be 98.7 (see section B.2).

First we calculate the ICC based on the CSMFs for Tanzania and k=0.25. For each cause we will have another ICC. Yet, for the calculations of the MIS, we would need a single ICC, which is appropriate for most causes.

	A	B	C	D
1	**Rank**	**CSMF [π]**	**ICC (=$B$23^2*(B/(1-B)))**	**Check (=IF(C2>$B$25;"false";"true"))**
2	1st	0.10950	0.00769	false
3	2nd	0.09982	0.00693	false
4	3rd	0.06890	0.00462	true
5	4th	0.06784	0.00455	true
6	5th	0.04123	0.00269	true
7	6th	0.04041	0.00263	true
8	7th	0.03766	0.00245	true
9	8th	0.03442	0.00223	true
10	9th	0.03014	0.00194	true
11	10th	0.02284	0.00146	true
12	11th	0.02229	0.00143	true
13	12th	0.01861	0.00119	true
14	13th	0.01732	0.00110	true
15	14th	0.01650	0.00105	true
16	15th	0.01275	0.00081	true
17	16th	0.01216	0.00077	true
18	17th	0.01180	0.00075	true
19	18th	0.01156	0.00073	true
20	19th	0.01120	0.00071	true
21	20th	0.01022	0.00065	true
22				
23	k	0.25		
24	Overall ICC	0.005		

You will need to fill the cells highlighted in yellow based on your CSMFs and the selected k. The formula used to calculate the ICC is given in A.5.2. In the cell B24 you can play around with a possible overall ICC to use in the MIS calculations. Column D will notify you with "false" if the calculated ICC in column C is bigger than the overall ICC. In our case we will decide for an **overall ICC of 0.005** as this covers almost all causes except for the first two. For the first two causes this mean, that selecting an ICC of 0.005 could potentially increase the minimal percentage change you would like to be able to detect (uncertainty range) from the first to the second three years. However, as you will see later on in B.5 the uncertainty range for detecting significant CSMF changes for the top two causes will be the smallest and thus being able to potentially measure only a slightly bigger change is not such a problem.

For calculating the coefficient of variation in cluster size, cv, you will now need your sampling frame. The cluster size in the case of CSMFs is the number of deaths per cluster. Thus, we are interested the variable "Estimated expected number of deaths in 2017". For this variable you will need to calculate its mean and the standard deviation. As reference it is also a good idea to look at the highest and the lowest value. Based on this we then calculate the cv and the MIS using an ICC of 0.005.

	A	B	C	D	E	F
1	**Ward**	**Estimated deaths (2017)**				
2	Mwamtani	58		Min (=MIN(B2:B3077))	6	
3	MtanilaWard	75		Max (=MAX(B2:B3077))	966	
4	MgungiraWard	39		Average (=AVERAGE(B2:B3077))	99	
5	Mkulazi	33		SD (=STDEV.S(B2:B3077))	87.5	
6	NambilajeWard	44		**cv (=E5/E4)**	0.887	
7	Orbomba	35		ICC	0.005	
8	RuvuWard	101		**MIS (=(1+((E6^2+1)*E4-1)*E7)/(1+(E4-1)*E7))**	1.26	
9	Liwale"B"Ward	62				
10	NsendaWard	111				
11	Chisano	40				
12	UkwegaWard	107				
13	IpoleWard	44				
14	KilaganoWard	88				
15	IssunaWard	73				
16	KigalaWard	26				
17	KilondoWard	19				
18	RuandaWard	54				
19	Ketumbeine	32				
20	Matuli	39				
21	Bukundi	59				
22	MaposeniWard	41				
23	MtundaWard	69				
24	UgallaWard	67				
25	ChibulaWard	21				
26	Machochwe	75				
27	Fufu	24				
28	Nondwa	104				
29	Kijungu	36				
30	IyumbuWard	56				
31	Sofi	103				

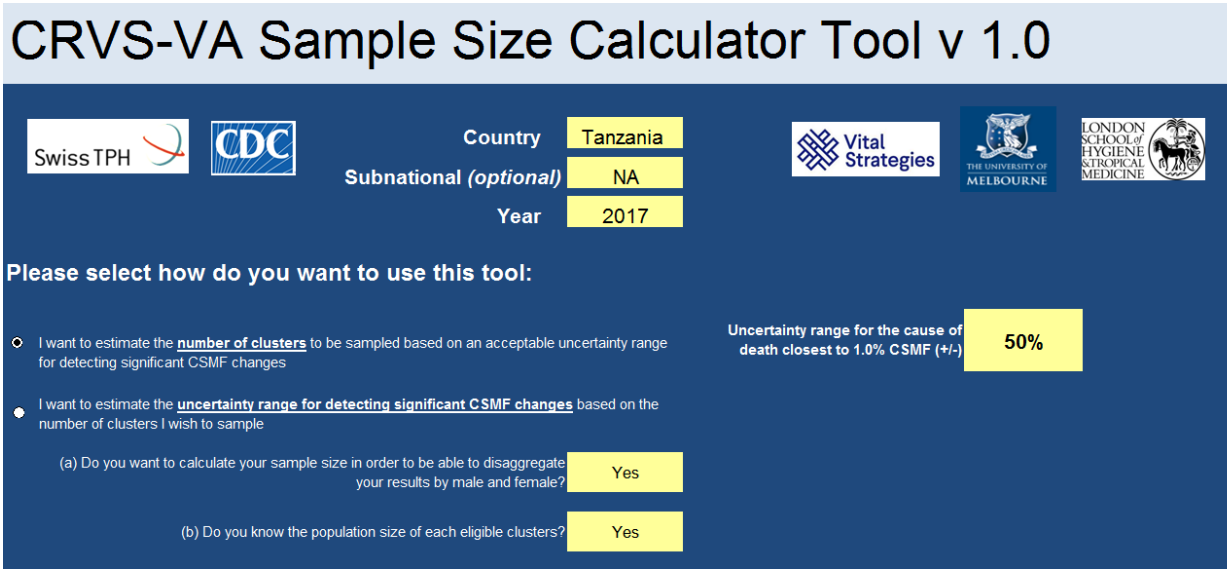
In the case of Tanzania this results in a **cv of 0.89 and a MIS of 1.26**. As you will see later on (B.5), an MIS of 1.5 would be needed to come up with a similar number of clusters required as if the harmonic mean was used all other parameters being constant.

B.5. Calculating the CRVS VA sample size

For calculating the sample size (required number of clusters) we will us the Sample Size Calculator Tool and the instructions given in Part B, section 3.2.

We will choose the option to estimate the number of clusters based on an acceptable uncertainty range for detecting significant CSMF changes. Therefore, we will need to enter the uncertainty range for detecting significant CSMF changes acceptable for us for the cause of death, which is closest to 1% CSMF in the top 20 list. In our case this is the 20th cause, which is "Interpersonal violence" with a CSMF of 1.2% (see section B.1.4). We will need to give an estimate what percentage change we would like to be able to detect for the CSMF of "Interpersonal violence" between the first three years and the second three years. For the moment, we will assume that 50% is acceptable for us. This means we think that for the CSMF of "Interpersonal violence" we would like to detect a decrease to 0.6% or less or an increase to 2.4% or more.

In the next step we will select "Yes", meaning we want our sample size to be calculated in a way that we can disaggregate our results by male and female. Also, given we prepared our sampling frame we will say "Yes" for the question whether we know the population size of each eligible cluster.

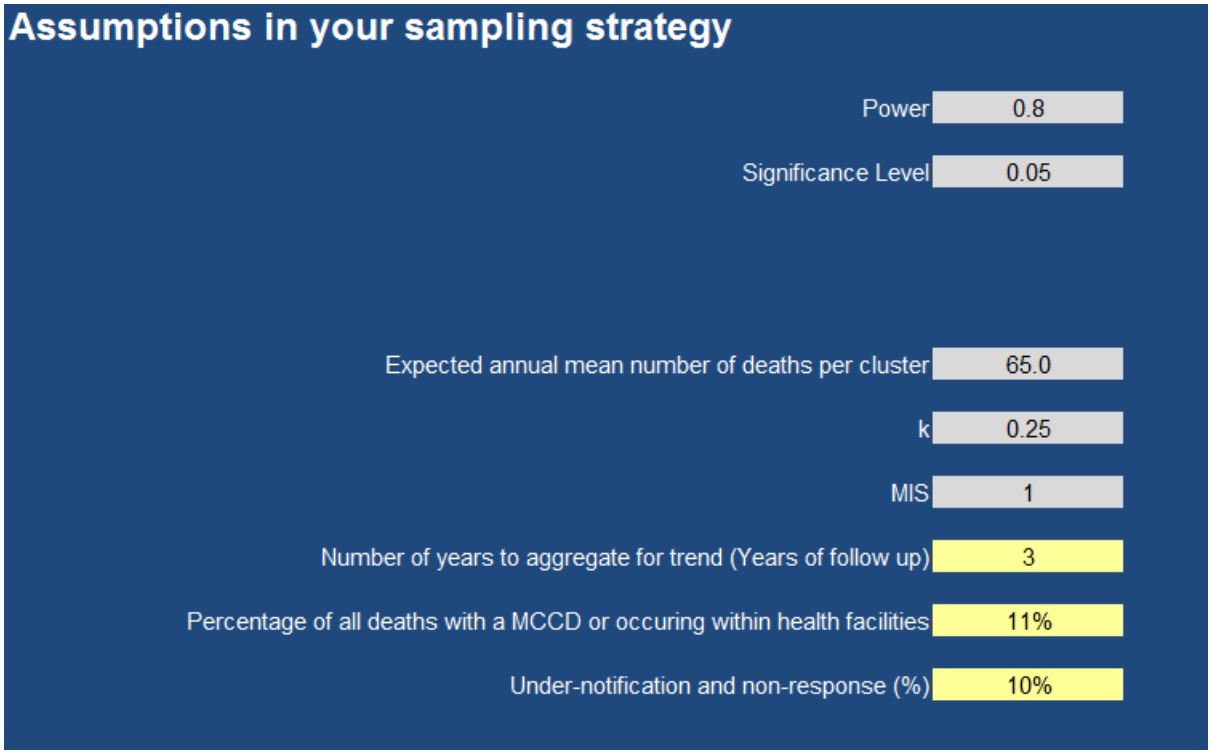


In a next step we will copy and paste the name, population size and crude death rate of each eligible cluster from the sampling frame into the provided table in the tool.

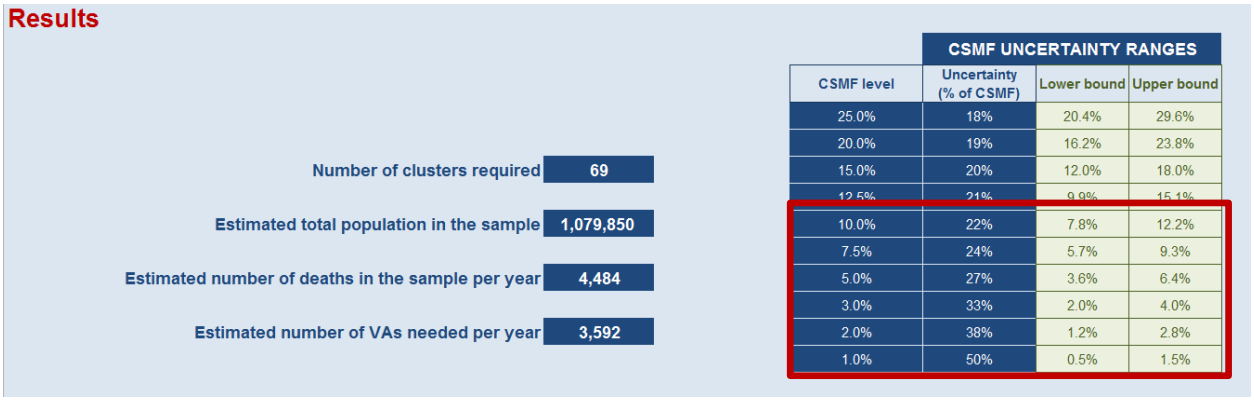
Name of the cluster unit	Cluster population size	Crude Death Rate (CDR) per 1,000	Expected annual number of deaths
UsukaWard	6,333	9.12	58
ItwangiWard	11,037	6.55	72
Minepa	19,090	6.22	119
KakunyuWard	23,158	7.70	178
StesheniWard	7,735	6.28	49
Kisesa	21,454	5.00	107
ChiolaWard	5,661	6.89	39
MbingaMhaluleWard	9,096	6.69	61
NandagalaWard	5,476	6.89	38
NyalumbuWard	15,851	8.45	134
Kibaigwa	27,640	6.08	168
BondelaSongweWard	20,228	7.23	146
NyakabangaWard	21,226	7.70	163
Chisano	6,468	6.22	40

This will automatically compute the expected annual number of deaths in each cluster from which the tool will subsequently calculate the expected annual mean number of deaths per cluster using the harmonic mean. To cross check the calculations, the automatically calculated expected annual mean number of deaths per cluster should be the same as calculated above from our sampling frame using the harmonic mean (B.2;64.97327).

Finally, based on our work in B.1 to B.3, our input parameters are as follows. For the power and significance level we assume the default values (**Power = 0.08 and significance level = 0.05**).



The results reveal the following:



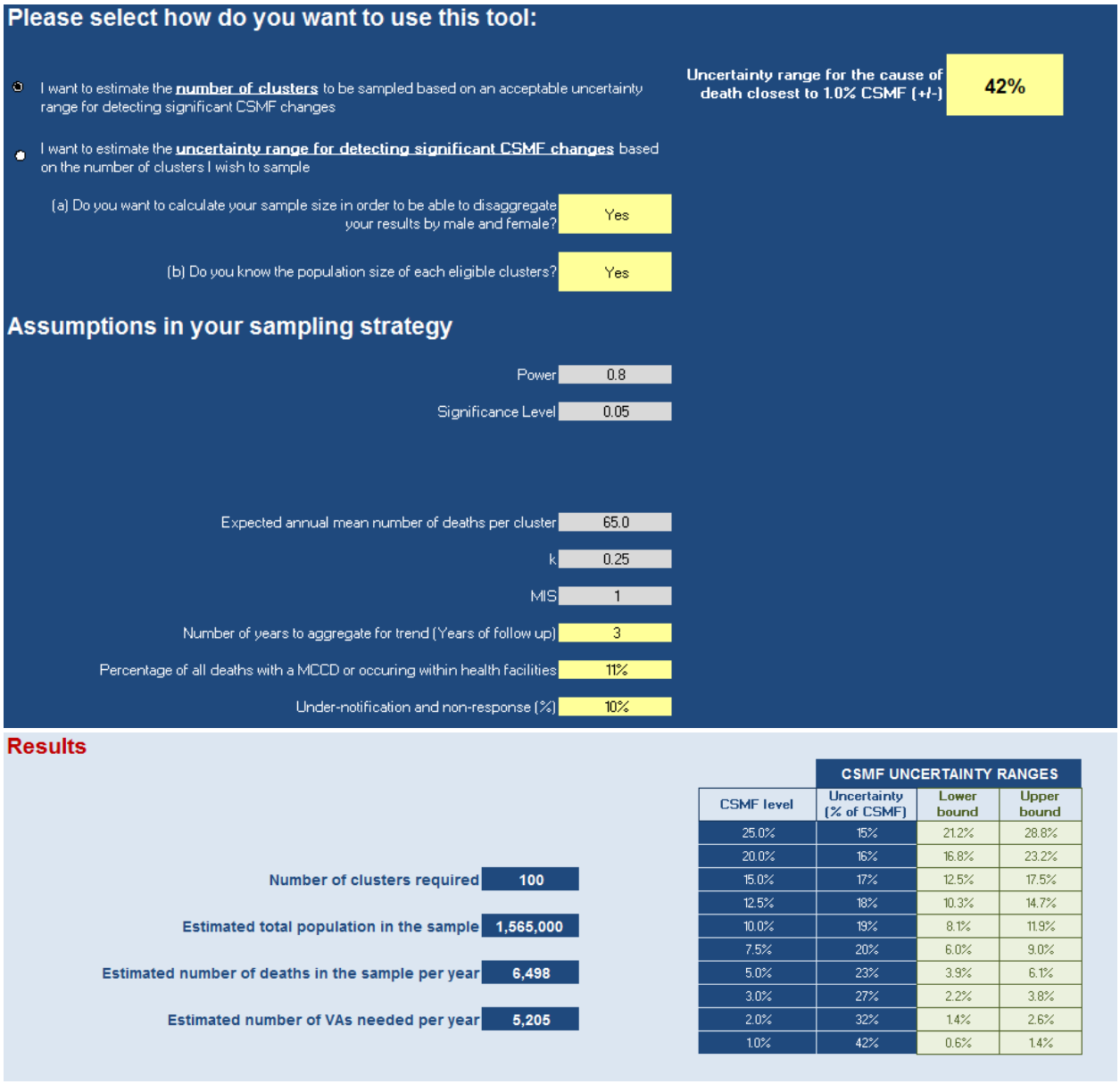
	CSMF UNCERTAINTY RANGES		
CSMF level	Uncertainty (% of CSMF)	Lower bound	Upper bound
25.0%	18%	20.4%	29.6%
20.0%	19%	16.2%	23.8%
15.0%	20%	12.0%	18.0%
12.5%	21%	9.9%	15.1%
10.0%	22%	7.8%	12.2%
7.5%	24%	5.7%	9.3%
5.0%	27%	3.6%	6.4%
3.0%	33%	2.0%	4.0%
2.0%	38%	1.2%	2.8%
1.0%	50%	0.5%	1.5%

For Tanzania, we know that the CSMFs range from 11% to 1%. Thus, we are in particular interested in the uncertainty ranges of CSMFs between 10% and 1%. For example, looking at our CSMF frequency distribution (section B.1.4), we see that "Lower respiratory infections" make up 10% of total deaths. This means that if between the first and the second three years the CSMF for "Lower respiratory infections" were to change to 7.8% or less or to 12.2% or more (plus/minus 22% or more), we would be able to detect this difference with for us acceptable statistical errors (power and significance level, also see A.2).

Note: The uncertainty range for detecting significant CSMF changes for the cause of death at the 1% CSMF level will always be larger than the uncertainty range in any other cause.

We can now play around with the uncertainty range of the cause of death closest to 1% ("Interpersonal violence") and see how this affects the number of clusters to be sampled and the uncertainty ranges for all other causes.

In the case of Tanzania, we decide that conducting VA in 100 clusters is on one hand feasible, (keeping in mind logistical and financial considerations), but also leads to a minimal detectable difference (uncertainty range) in the CSMF for the 20th cause (cause of death closest to 1%) of 42.3%, which is satisfactory and acceptable for us. The final input parameters are the following:



Please select how do you want to use this tool:

- I want to estimate the **number of clusters** to be sampled based on an acceptable uncertainty range for detecting significant CSMF changes
- I want to estimate the **uncertainty range for detecting significant CSMF changes** based on the number of clusters I wish to sample

Uncertainty range for the cause of death closest to 1.0% CSMF (+/-): 42%

(a) Do you want to calculate your sample size in order to be able to disaggregate your results by male and female? Yes

(b) Do you know the population size of each eligible clusters? Yes

Assumptions in your sampling strategy

Power	0.8
Significance Level	0.05
Expected annual mean number of deaths per cluster	65.0
k	0.25
MIS	1
Number of years to aggregate for trend (Years of follow up)	3
Percentage of all deaths with a MCCD or occuring within health facilities	11%
Under-notification and non-response (%)	10%

Results

Number of clusters required	100
Estimated total population in the sample	1,565,000
Estimated number of deaths in the sample per year	6,498
Estimated number of VAs needed per year	5,205

	CSMF UNCERTAINTY RANGES		
CSMF level	Uncertainty (% of CSMF)	Lower bound	Upper bound
25.0%	15%	21.2%	28.8%
20.0%	16%	16.8%	23.2%
15.0%	17%	12.5%	17.5%
12.5%	18%	10.3%	14.7%
10.0%	19%	8.1%	11.9%
7.5%	20%	6.0%	9.0%
5.0%	23%	3.9%	6.1%
3.0%	27%	2.2%	3.8%
2.0%	32%	1.4%	2.6%
1.0%	42%	0.6%	1.4%

Note: You could now also select “No” for the question whether you know the population size of each eligible cluster and enter mean population per cluster (15,650) and the crude death rate (6.351) based on the inputs from B.2. With an MIS of 1.5 you would then end up with a similar number of clusters required (104) as if the harmonic mean was used all other parameters being constant.

B.6. Sampling strategy for selecting the CRVS VA sample clusters

Based on the calculations done in B.5 we now need to sample 100 wards from our sampling frame. Given that we want our sample to reflect characteristics of the whole country, we decide to stratify by region and urban/rural status. To do so we follow the instructions given in Part B, section 4.1 and obtain the number of wards (clusters) we need to sample per strata:

	A	B	C	D	E	F	G
1	**STRATA**	**Rural**			**Urban**		
2	**Region**	**Number of Wards in 2012**	**Expected population in 2017**	**Wards required (=C/$E$30*$E$32)**	**Number of Wards in 2012**	**Expected population in 2017**	**Wards required (=F/$E$30*$E$32)**
3	Arusha	86	1,285,137	3	19	481,423	1
4	DaresSalaam	0	0	-	90	5,786,614	12
5	Dodoma	145	1,785,026	4	37	463,352	1
6	Geita	96	1,960,259	4	0	0	-
7	Iringa	60	649,193	1	19	244,965	1
8	Kagera	166	2,760,905	6	14	162,915	0
9	Katavi	21	317,111	1	9	126,947	0
10	Kigoma	80	1,886,919	4	28	480,988	1
11	Kilimanjaro	127	1,553,630	3	21	208,827	0
12	Lindi	88	660,648	1	18	109,116	0
13	Manyara	98	1,413,949	3	8	97,882	0
14	Mara	138	1,790,055	4	13	149,945	0
15	Mbeya	169	2,414,000	5	37	597,260	1
16	Morogoro	121	2,000,156	4	19	372,524	1
17	Mtwara	118	1,089,863	2	27	228,012	0
18	Mwanza	133	2,366,528	5	21	869,115	2
19	Njombe	61	423,038	1	19	234,099	0
20	Pwani	86	978,014	2	11	164,982	0
21	Rukwa	48	945,274	2	15	251,163	1
22	Ruvuma	91	1,010,817	2	21	253,346	1
23	Shinyanga	81	1,235,063	3	37	473,910	1
24	Simiyu	107	1,661,813	3	0	0	-
25	Singida	97	1,296,905	3	16	172,112	0
26	Tabora	127	2,217,285	5	25	249,621	1
27	Tanga	176	1,891,588	4	32	368,025	1
28	**TOTAL**	**2,520**	**35,593,176**	**74**	**556**	**12,547,143**	**26**
29							
30	**Total population (=C28+F28)**				48,140,318		
31	**Total umber of wards (clusters) (=B28+E28)**				3,076		
32	**Number of wards to be sampled:**				100		

Cells highlighted in yellow require input from your sampling frame or the above sample size calculations.

To now pick the number of required wards from each stratum, we will use PPS sampling (see Part B, section 4.2). For this, we will need a list of all eligible wards per stratum. Here, we will work out the example of Arusha urban, where we pick one ward from 19 wards. You would need to do this for each individual stratum.

To implement PPS sampling we will follow the steps in Part B, section 4.2 and add a couple of Excel specific features to facilitate the work.

1) In column A to F we list characteristics of all clusters in the stratum Arusha urban. This information comes from the sampling frame.
2) In column G we calculate the cumulative sum of the stratum population. This means row 2 equals F2, row 3 equals G2 + F3 and this formula will then need to be dragged down for all subsequent rows. G20 is the total population size in the stratum (also in E22).

3) In E24 we calculate the sampling interval by dividing the total population in the stratum by the number of required clusters in the stratum. The number of clusters to be sampled in the stratum comes from above and is 1 in the case of Arusha urban.
4) In E25 we choose a random number between 0 and 1. The random start shown in E26 is then given by this number multiplied by the sampling interval (E24).
5) To automate the calculation of the series: RS; RS+SI; RS+2*SI; … RS+(d-1)*SI, we will use some intermediate steps in Excel:
    a. In column H and I we define the lower and the upper bound of the cumulative population in each cluster. In row 2 this is 0 (column H) to G2 (column I) and in row 3 this is I2 (column H) to G3 (column I). The formulas from row 3 you will then need to drag down till the end of your list.
    b. In column J we define the target which is the value of the series, meaning the random start plus the sampling interval times column L (number of wards already selected). Again the formula in row 2 is different from the rest and the formula given for row 3 will need to be dragged down to the last cluster in the list.
    c. Column K indicates with 1 if a ward is selected and with 0 if the ward is not selected. Here the formula for row 2 can be entered and then dragged down to the last cluster.
    d. Column L sums up how many wards were already selected and is primarily needed to calculate the target in column J. Again, the formula in row 2 is different from the rest and the formula given for row 3 will need to be dragged down to the last cluster in the list.
6) The clusters selected are those with a 1 in column K. In our case this would be the ward called “Sombetini”.

Note: Every time you refresh the Excel file, the random number in E25 will be different and therewith the clusters selected will be different.

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Zone**	**Region**	**District**	**Ward**	**Unique Ward ID**	**Estimated population size (2017)**	**Cumulative population (row2=F2; row3=G2+F3)**	**Lower Bound (row2=0; row3=I2)**	**Upper Bound (=G2)**	**Target (row2=$E$26; row3=$J$2+$E$24*L2)**	**Selected Ward (=IF(AND(J2>H2; J2<=I2);1;0))**	**Cumulative selected ward (row2=K2; row3=K3+L2)**
2	Northern	Arusha	ArushaUrban	Kaloleni	10020792	11,088	11,087.6	0.0	11,087.6	68,532.7	0	0
3	Northern	Arusha	ArushaUrban	Themi	10022986	10,934	22,021.4	11,087.6	22,021.4	68,532.7	0	0
4	Northern	Arusha	ArushaUrban	Sombetini	10022925	55,800	77,821.1	22,021.4	77,821.1	68,532.7	1	1
5	Northern	Arusha	ArushaUrban	Kati	10020879	3,600	81,421.0	77,821.1	81,421.0	549,956.2	0	1
6	Northern	Arusha	ArushaUrban	Kimandolu	10021054	31,963	113,384.4	81,421.0	113,384.4	549,956.2	0	1
7	Northern	Arusha	ArushaUrban	Moshono	10022018	23,928	137,312.1	113,384.4	137,312.1	549,956.2	0	1
8	Northern	Arusha	ArushaUrban	Engutoto	10020363	8,585	145,896.8	137,312.1	145,896.8	549,956.2	0	1
9	Northern	Arusha	ArushaUrban	SokonI	10022918	84,774	230,670.4	145,896.8	230,670.4	549,956.2	0	1
10	Northern	Arusha	ArushaUrban	UngaLtd	10023057	20,048	250,718.4	230,670.4	250,718.4	549,956.2	0	1
11	Northern	Arusha	ArushaUrban	Lemara	10021299	22,617	273,335.2	250,718.4	273,335.2	549,956.2	0	1
12	Northern	Arusha	ArushaUrban	Ngarenaro	10022455	14,314	287,649.2	273,335.2	287,649.2	549,956.2	0	1
13	Northern	Arusha	ArushaUrban	Oloirien	10022698	21,594	309,242.9	287,649.2	309,242.9	549,956.2	0	1
14	Northern	Arusha	ArushaUrban	Elerai	10020345	47,107	356,350.4	309,242.9	356,350.4	549,956.2	0	1
15	Northern	Arusha	ArushaUrban	Baraa	10020030	14,448	370,798.5	356,350.4	370,798.5	549,956.2	0	1
16	Northern	Arusha	ArushaUrban	Daraja2	10020317	22,532	393,330.9	370,798.5	393,330.9	549,956.2	0	1
17	Northern	Arusha	ArushaUrban	Terrat	10022984	25,190	418,521.0	393,330.9	418,521.0	549,956.2	0	1
18	Northern	Arusha	ArushaUrban	Olasiti	10022680	42,035	460,555.8	418,521.0	460,555.8	549,956.2	0	1
19	Northern	Arusha	ArushaUrban	Sekei	10022855	10,651	471,206.4	460,555.8	471,206.4	549,956.2	0	1
20	Northern	Arusha	ArushaUrban	Levolos	10021304	10,217	481,423.4	471,206.4	481,423.4	549,956.2	0	1
21												

22	**Total statrum population (=N20)**	481,423
23	**Number of clusters to be sampled**	1
24	**Sampling interval (=E22/E23)**	481423.43
25	**Random between 0 and 1 (=RAND())**	0.1423544
26	**Random start (=E25*E24)**	68533

B.7. Calculating the number of clusters required in order to disaggregate results for male and female

As explained in A10.1 the Sample Size Calculator Tool doubles the calculated number of clusters required in order to have disaggregated results for male and female. It is also stated, that this however is only correct if the CDRs for female and male are somewhat similar.

To investigate if this is the case in Tanzania, we will calculate the required number of clusters for the sub-population of the sex with the lower CDR, which is the female population.

In a first step we will need to compute the **estimated female population for each cluster** by simply dividing the total population by two (assuming you have 50% male and 50% female in the population). We also need to compute **the regional female CDR for 2017** using the above estimated national female CDR of 5.81 (section B.1.4).

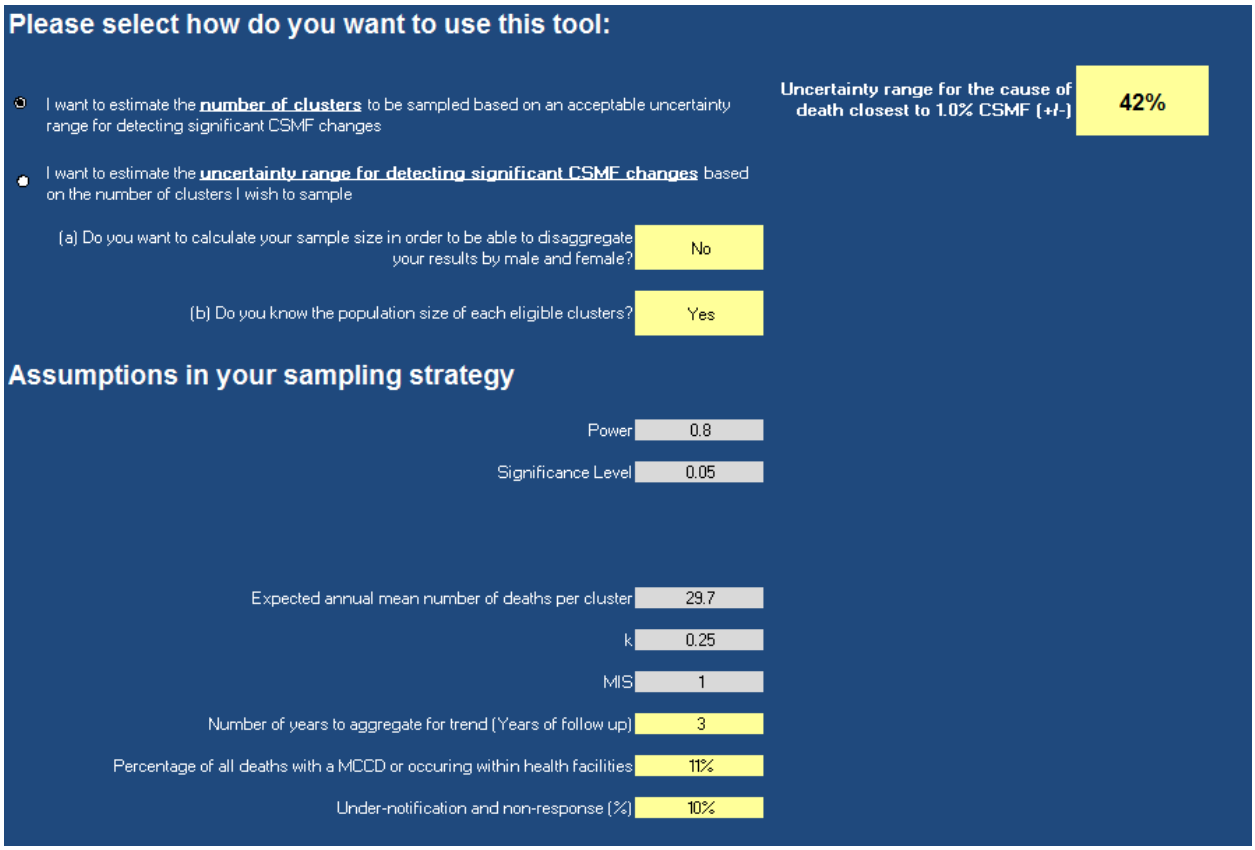
Note: Alternatively, if the regional CDRs do not vary a lot, you could also simply divide the mean population per cluster by two (assuming you have 50% male and 50% female in the population) and multiply this by the national female CDR/1,000 (15'650/2*5.81/1,000 = 45.47).

In a next step we go back to our Sample Size Calculator Tool for which we used the final input parameters listed in B.5 and which revealed that we would need to sample 100 clusters if we wanted to be able to measure a percentage change of 42.3% or more in CSMF for the 20th cause of death. We will now replace the cluster population size and the crude death rate in our table to the female population and its specific crude death rate.

Name of the cluster unit	Cluster population size	Crude Death Rate (CDR) per 1,000	Expected annual number of deaths
UsukaWard	3,167	8.34	26
ItwangiWard	5,519	6.00	33
Minepa	9,545	5.69	54
KakunyuWard	11,579	7.05	82
StesheniWard	3,867	5.75	22
Kisesa	10,727	4.57	49
ChiolaWard	2,831	6.30	18
MbingaMhaluleWard	4,548	6.12	28
NandagalaWard	2,738	6.30	17
NyalumbuWard	7,926	7.73	61
Kibaigwa	13,820	5.56	77
BondelaSongweWard	10,114	6.61	67
NyakabangaWard	10,613	7.05	75
Chisano	3,234	5.69	18

The expected annual mean number of deaths per cluster changes from 65.0 to 29.7.

The only thing which we now change in addition is to select "No" for the question whether we want our sample size to be calculated in a way that we can disaggregate our results by male and female. We then look at the results and see if the required number of clusters increased or decreased form what we had before.



Results

Number of clusters required	102
Estimated total population in the sample	1,596,300
Estimated number of deaths in the sample per year	3,032
Estimated number of VAs needed per year	2,429

	CSMF UNCERTAINTY RANGES		
CSMF level	Uncertainty (% of CSMF)	Lower bound	Upper bound
25.0%	12%	21.9%	28.1%
20.0%	13%	17.4%	22.6%
15.0%	14%	12.9%	17.1%
12.5%	15%	10.6%	14.4%
10.0%	16%	8.4%	11.6%
7.5%	18%	6.1%	8.9%
5.0%	21%	3.9%	6.1%
3.0%	26%	2.2%	3.8%
2.0%	31%	1.4%	2.6%
1.0%	42%	0.6%	1.4%

In our case the number of required clusters increased from 100 to 102. This means our original sample size calculations, which simply doubled the number of required clusters, was slightly to low. However, the difference is so small that we decide to remain with the simpler calculation.

If the number of clusters needed for “females only” would be much bigger than the number of clusters needed based on your original sample size calculations, then simply doubling the numbers of clusters required (as done by the Sample Size Calculator Tool) is insufficient. In this case the difference between the CDR of female and male is so big that you would need to calculate your sample size for the sub-population of the sex with the lower CDR only.

B.8. Calculating the number of clusters required separately for areas with different proportions of deaths with an MCCD

For the above example of Tanzania the percentage of total deaths with a MCCD was assumed to be 11% nationally (see B.1.4). However, this assumption only holds true if the proportion of deaths with an MCCD is similar across the country. In Tanzania, we suspect that this could vary depending on the urban/rural status. Thus, we will investigate this further.

Based on DHIS2 data we know that 59% of all deaths with a MCCD come from health facilities in rural area and 41% from health facilities in urban areas (average of 2015 and 2016 data). From the calculations done in B.5 we know we need to sample 100 wards with a total estimated number of 9,940 deaths. In B.6 we calculated that 74 of the 100 clusters should be rural and 26 urban. With an expected annual mean number of deaths per cluster of 99.4, this would mean that 7,356 deaths in the sample are expected to be in rural areas and the rest in urban areas. This, together with the known national percentage of total deaths with a MCCD (11%) allows us to calculate the percentage of deaths with a MCCD in rural and urban areas.

	A	B	C	D	E	F	G
1		Expected deaths in the sample		% of deaths with MCCD	Expected deaths with MCCD in the sample		
2		#	%	%	#	%	
3	Rural	7,355.60	74%	9%	642.37	59%	
4	Urban	2,584.40	26%	17%	451.03	41%	
5	**Total**	**9,940.00**	**100%**	**11%**	**1,093.40**	**100%**	
6							
7							

The input data is highlighted in yellow. In a first step you calculate the total number of expected deaths with a MCCD in the sample. In a second step you compute the expected number of deaths with a MCCD in the sample for rural and urban areas based on the percentage of MCCD deaths coming from either of these two areas. In a third step you then use the expected number of deaths in the sample and the expected number of deaths with a MCCD in the sample per rural/urban status to calculate the percentage of deaths with a MCCD in rural and urban areas. In our case this reveals that 9% of deaths in rural and 17% of deaths in urban areas have a MCCD.

All other parameters being constant and the same in urban and rural areas, this will now impact how many clusters we have to sample from each area. Thus, in the following steps we will use the urban/rural specific percentage of deaths with a MCCD and recalculate the numbers of cluster needed per area.

	A	B	C	D	E
1	**Steps**		**Rural**	**Urban**	**Total**
2		Number of clusters required (based on national estimate of 11% deaths with MMCD)	74.00	26.00	100.00
3	1.	Number of deaths in the sample	7,356	2,584	9,940
4	2.	Number of deaths to be included (minus 10% under-notification/non-response rate and 11% deaths with MCCD)	5,892	2,070	7,962
5					
6		% deaths with MCCD by rural/urban	9%	17%	11%
7	3.	Number of deaths based on rural/urban specific % deaths with MCCD (minus 10% under-notification/non-response rate)	6,042	1,920	7,962
8	4.	Revised number of clusters required (based on rural/urban specific % deaths with MMCD)	72.16	28.03	100.19
9					

The input data is highlighted in yellow, whereas the clusters needed and the total number of deaths in the sample come from B.5 and B.6, the under-notification and non-response rate (assumed to be constant across the country) and the national estimate of percentage of total deaths with a MCCD from B.1.4, and proportion of all deaths with a MCCD per urban/rural area from above.

1) Calculate the number of deaths in the sample for rural and urban areas based on expected annual mean number of deaths per cluster of 99.4 (B.5) (Rural: E3/(C2+D2)*C2; Urban: E3/(C2+D2)*D2).
2) Calculate the number of death to be included in the analysis by subtracting 10% for non-response and under notification and 11% for deaths with a MCCD (Rural: C3*(1-0.11)*(1-0.1); for Urban and Total replace C3 by D3 and E3).
3) Calculate the number of deaths based on the rural/urban specific percentage of deaths with a MCCD by subtracting 10% for non-response and under notification and the percentage of deaths with a MCCD (Rural: C3*(1-C6)*(1-0.1); for Urban and Total replace C with D and E).
4) Calculate the revised number of clusters required based on the rural/urban specific percentage of deaths with a MCCD (Rural: C2/C7*C4; for Urban and Total replace C by D and E).

After rounding, this reveals that we would need instead of 74 and 26 clusters 72 and 28 clusters in rural and urban areas, respectively. However, in the case of Tanzania this difference is very minor and thus we decide to remain with the simpler calculation and the assumption of a single national percentage of total deaths with a MCCD of 11% and 76 and 24 clusters in rural and urban areas.

Annex C. Additional Resource Material

C.1. National Crude Death Rate estimates for Data for Health Initiative CRVS VA countries

CDR [/1,000]	Source			
Country	World Population Review (2017)	Knoema (2017)	UN/population division (2015-2020)	World Bank (2015)
Bangladesh	5.265	5.27	5.271	5.31
Brazil	6.266	6.24	6.296	6.092
Colombia	6.057	6.08	6.122	5.942
Ecuador	5.104	5.12	5.118	5.127
Ghana	8.552	7.99	7.904	8.314
Indonesia	7.198	7.16	7.184	7.096
Kenya	7.641	5.66	5.681	5.841
Malawi	6.872	7.10	7.09	7.498
Morocco	5.67	5.12	5.124	5.145
Myanmar	8.333	8.19	8.236	8.101
Philippines	6.81	6.55	6.546	6.496
Papua New Guinea	7.6	7.09	7.089	7.133
Rwanda	6.43	5.84	5.814	6.132
Solomon Islands	5.567	4.70	4.68	4.852
Sri Lanka	7.031	6.99	7.044	6.814
Tanzania	6.351	6.49	6.351	7.015
Zambia	8.092	7.60	7.59	7.998

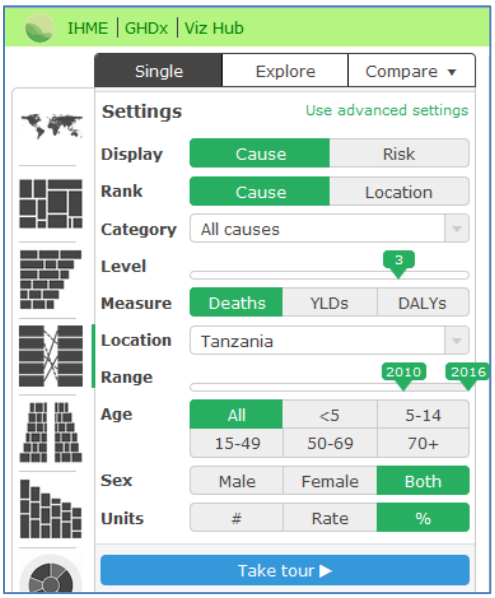
C.2. Method for estimating national level CSMFs

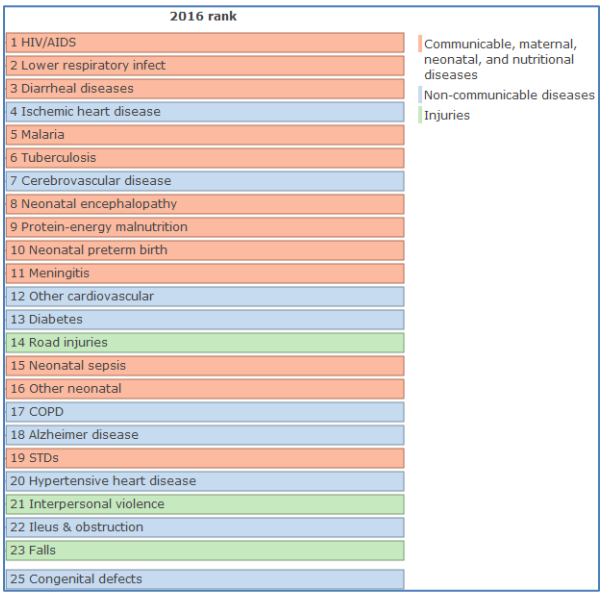
It is important to have an estimate of the CSMFs for the top 20 male and top 20 female causes at national level that are produced by your VA system. If such estimates are not available from prior studies, these are available on-line at https://vizhub.healthdata.org/gbd-compare/. This is the IHME GBD Compare web site.

1. Make the following settings:
   - Choose: Arrow Diagram
   - Set Tab to: **Single**
   - Set Display to: **Cause**
   - Set Rank to: **Cause**
   - Set Category to: **All causes**
   - Set Aggregation Level[11] to: **3**
   - Set Measure to: **Deaths**
   - Set Location to: **Your Country Name**
   - Range: any year till the year you are interested in (e.g.2010 to 2016)
   - Set Age to: **All**
   - Set Sex to your choice: **Male, Female, or Both**
   - Set Units to: **%**

Note: The same could be done for rates if this is also of interest for you.

Here is a example settings for Tanzania and associated output for 2016:

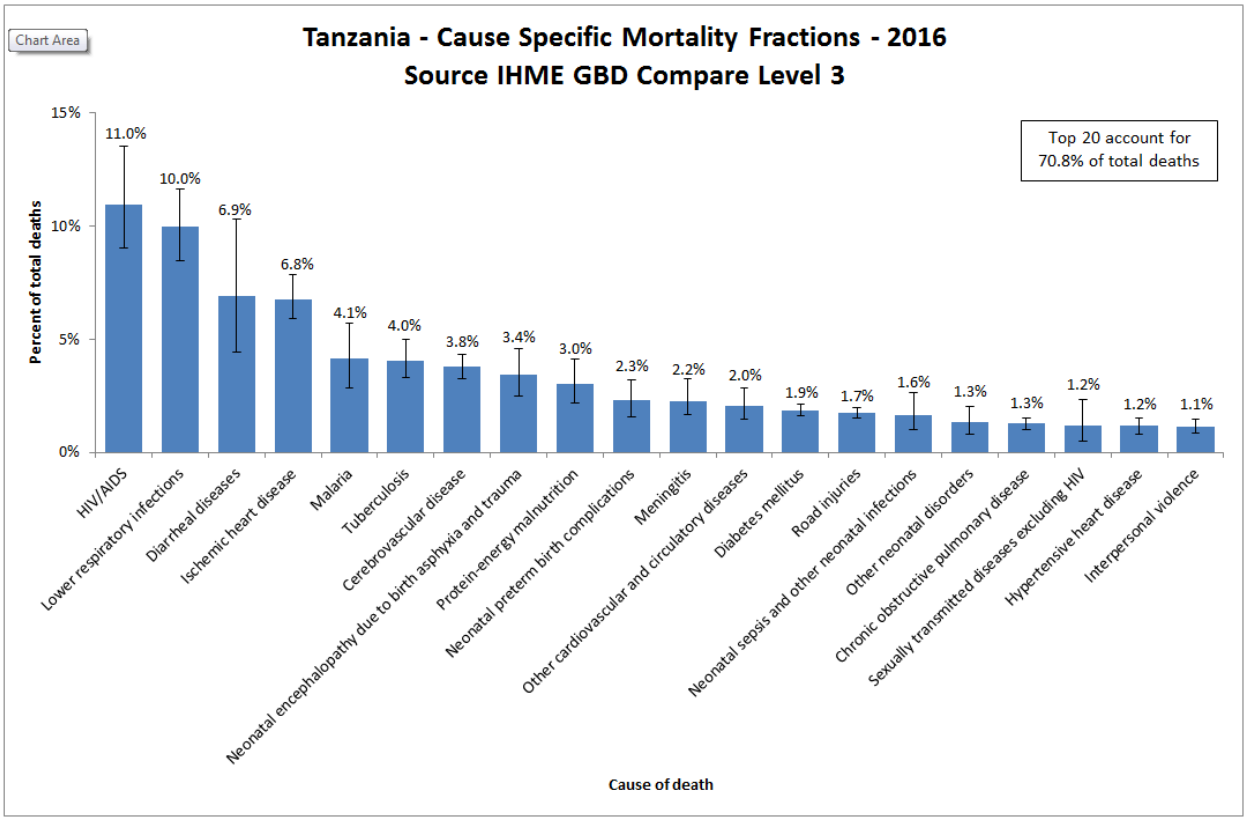




2. With the above settings still active, click on the Download button at the upper right and save the data as a csv file to open in MS Excel.

[11] Aggregation level 1 (Communicable, Non-communicable and Injury) is too coarse, and Aggregation Level 4 is too disaggregated to highly specific causes. Level 3 concords best with the VA target causes lists in general use today.

3. In Excel go to Data > Get External Data > From Text and import the csv file (delimiters is “comma”). This file has all proportional deaths for all causes and the selected years.
4. Select the most recent year and the measure “Percent of Total Deaths”.
5. Then sort on the value to get a ranked list of the CSMFs.
6. Delete causes that are not on your VA Target Cause list (e.g., Alzheimer disease and other dementias) to obtain the top 20 estimated CSMFs that will be seen in your VA data (best done by a clinician).
7. Afterwards, prepare a frequency distribution graphic, which will look something like this, which is an example from Tanzania for the top 20 IHME estimated CSMFs selected related to WHO 2016 VA target cause lists for 2016.



8. Note the range of CSMFs between the 1st and 20th cause. The distribution of CSMFs for your country will likely be such that the 1st rank will fall between 10 and 15%, while the 10th rank will be around 2%, and the 20th rank will be around 1%. The top 20 causes will likely account for about 70% of all deaths.

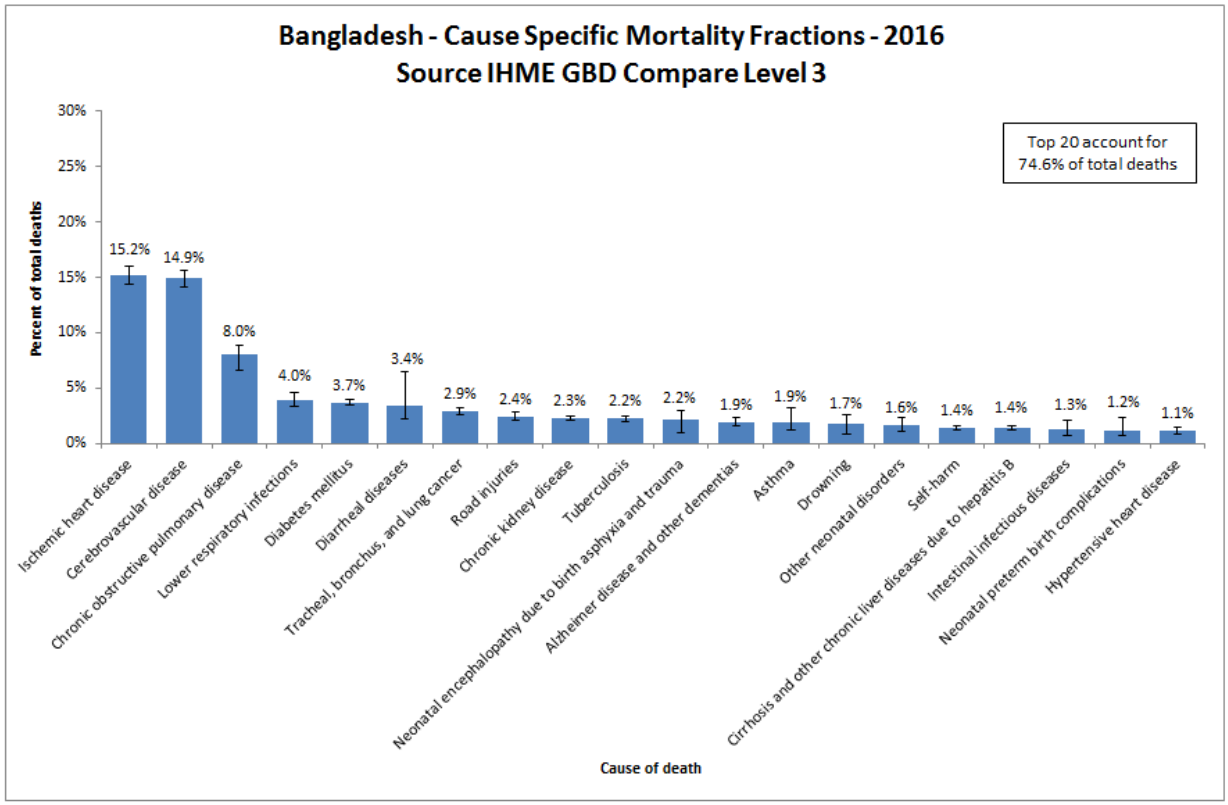
Estimated distributions for all countries participating in the Data for Health Initiative for 2016 are provided in Annex C.3.

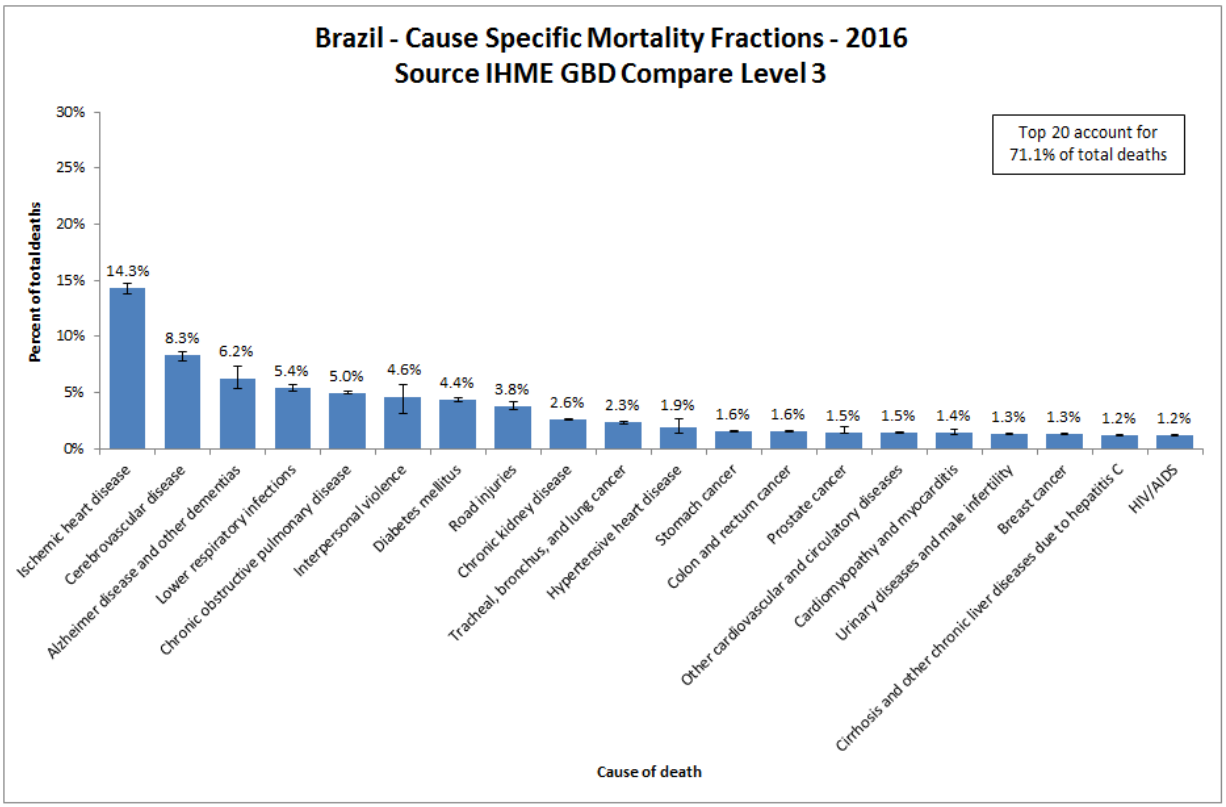
The current VA target cause lists for the WHO 2016 Standard VA[12] and for the SmartVA[13] methods are provided in C.4for comparison to your country’s distribution of leading causes.

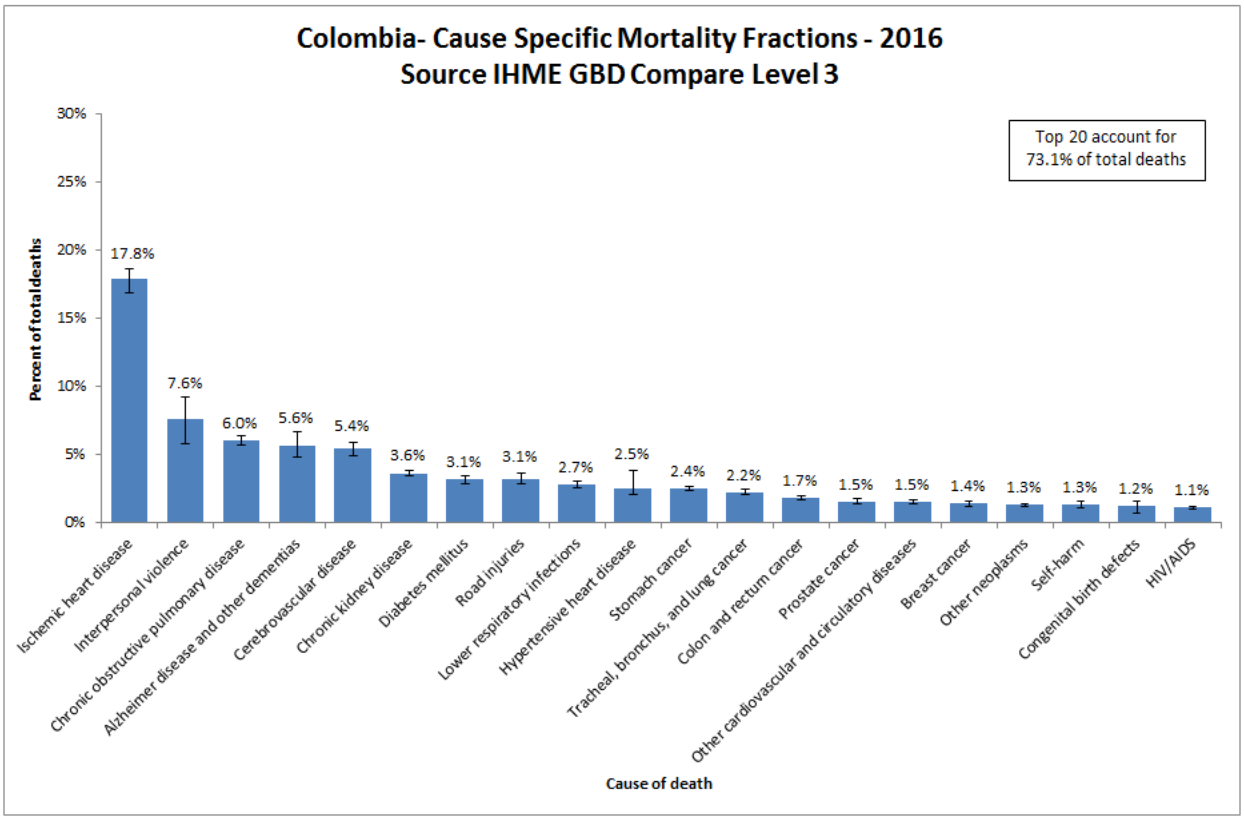
[12] Available at [insert URL]

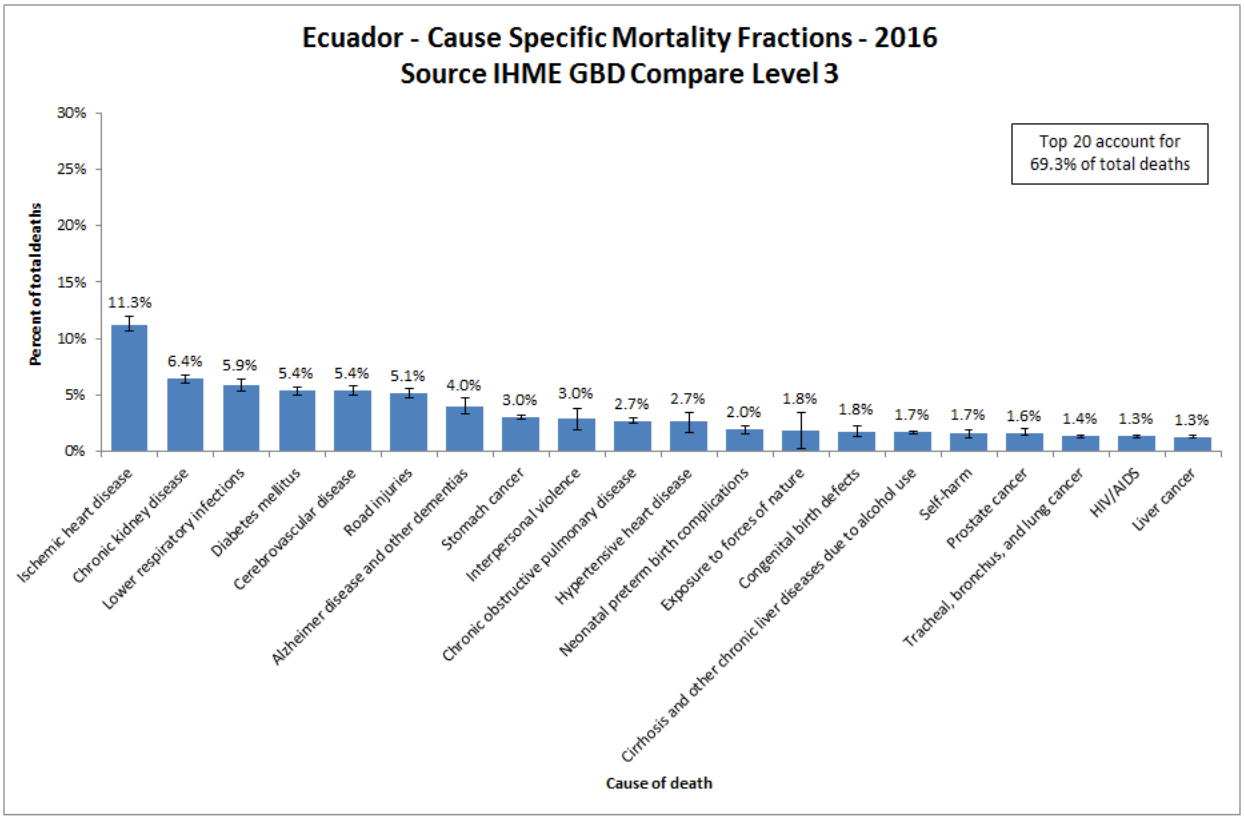
[13] Available at [insert URL]

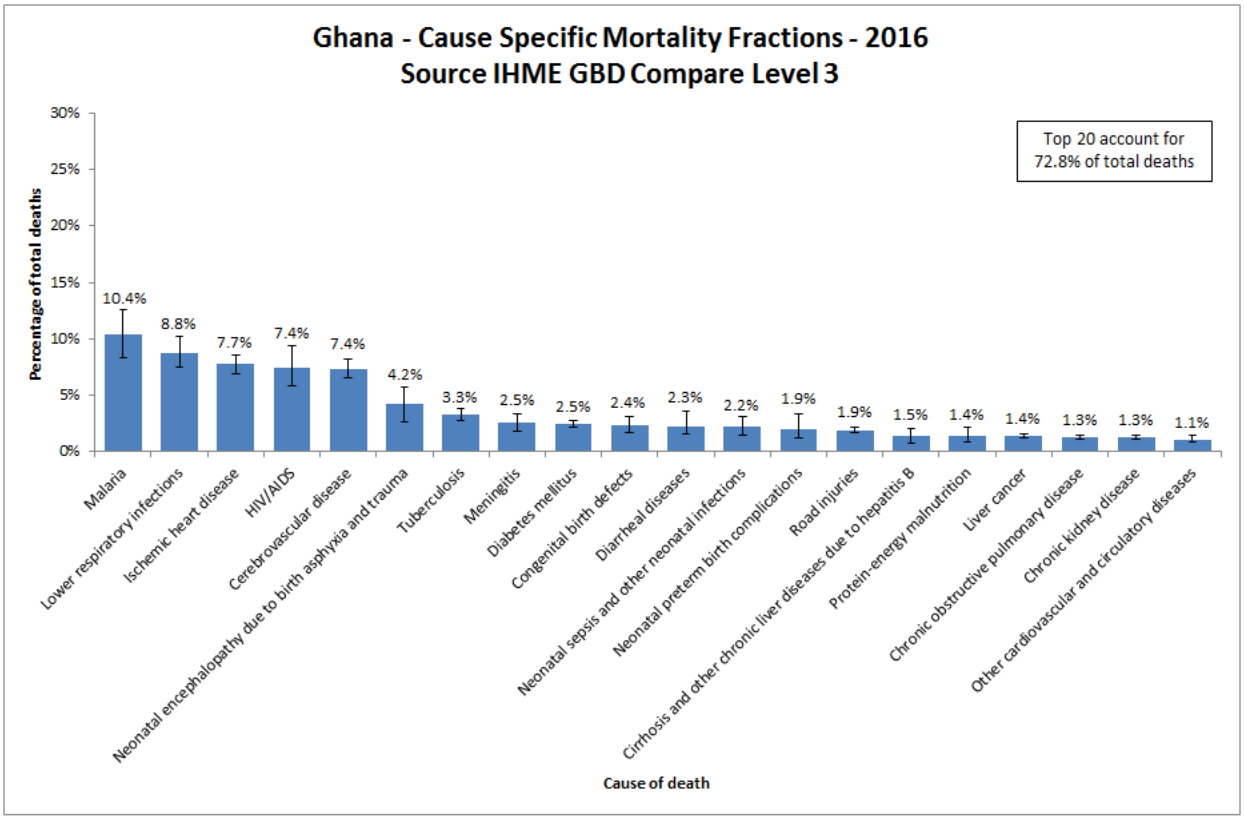
C.3. Top 20 CSMF estimates for Data for Health Initiative CRVS VA countries

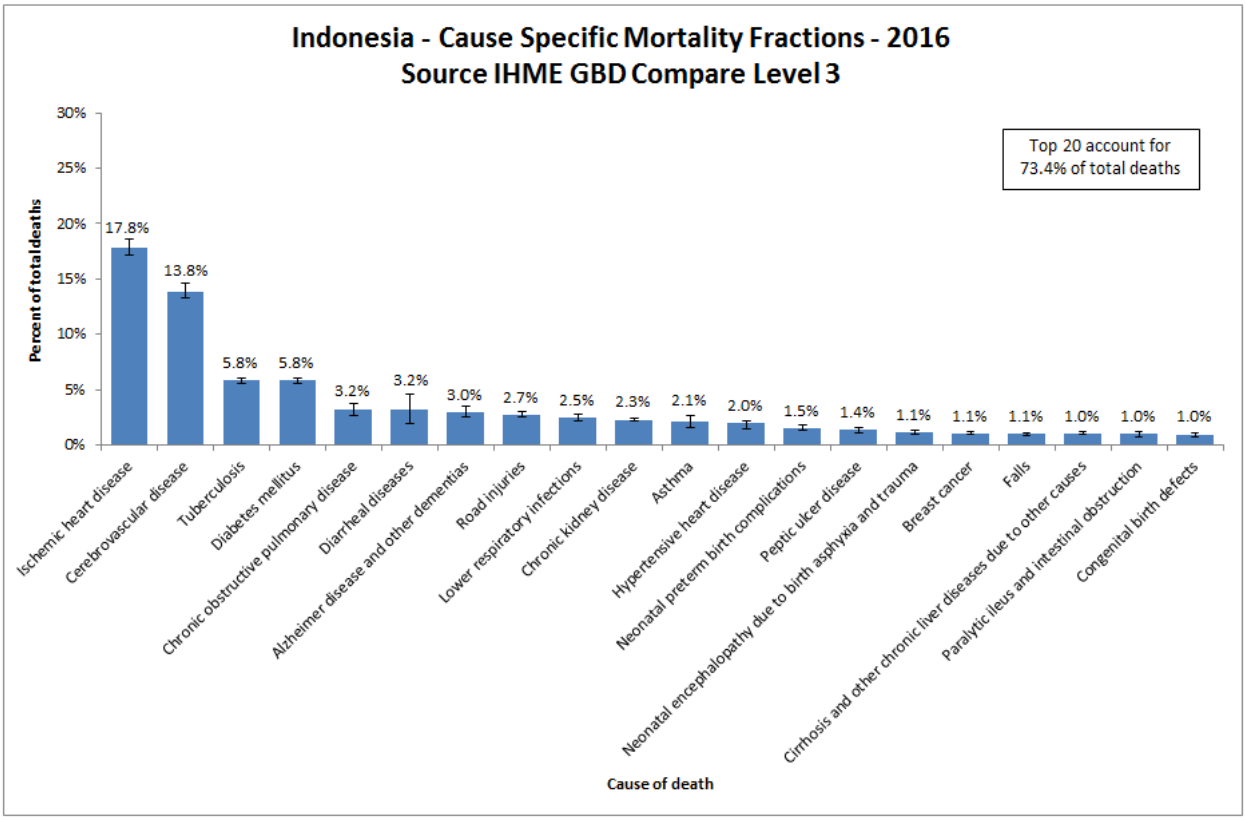


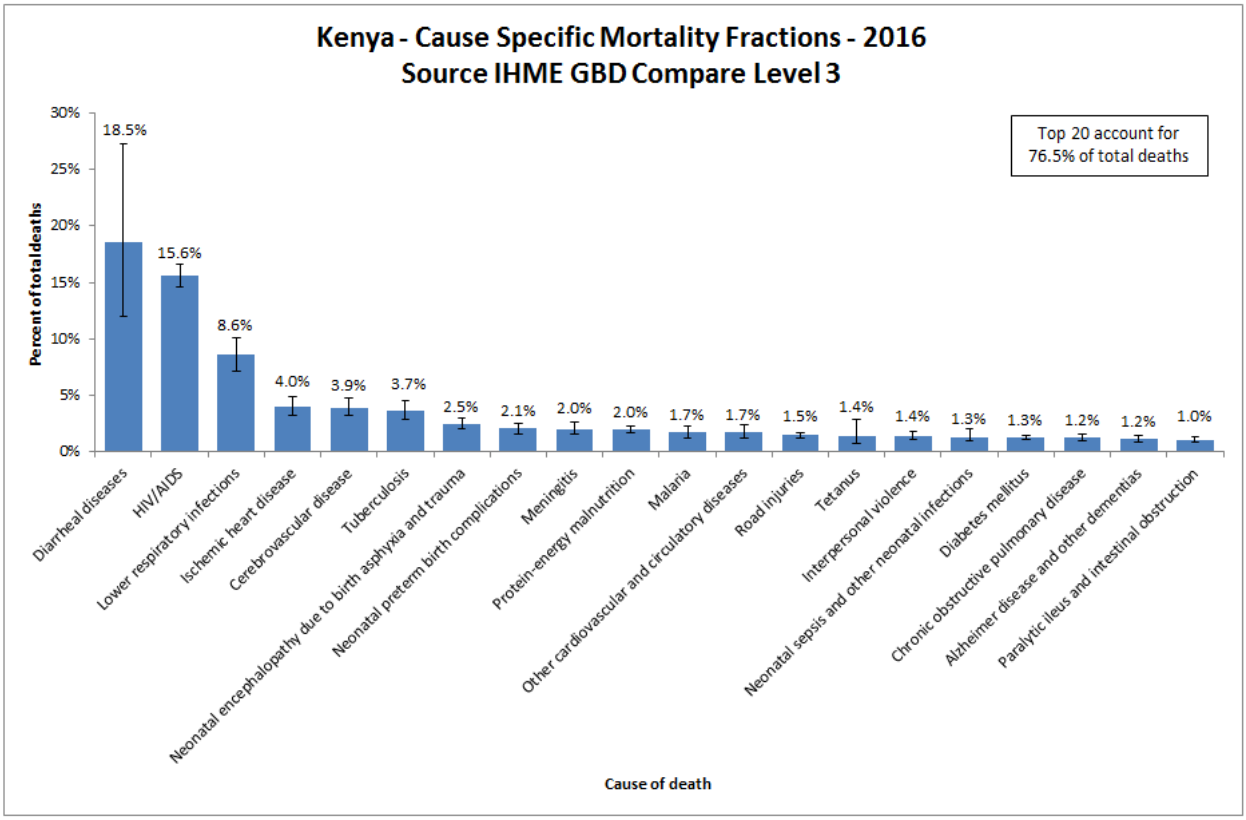


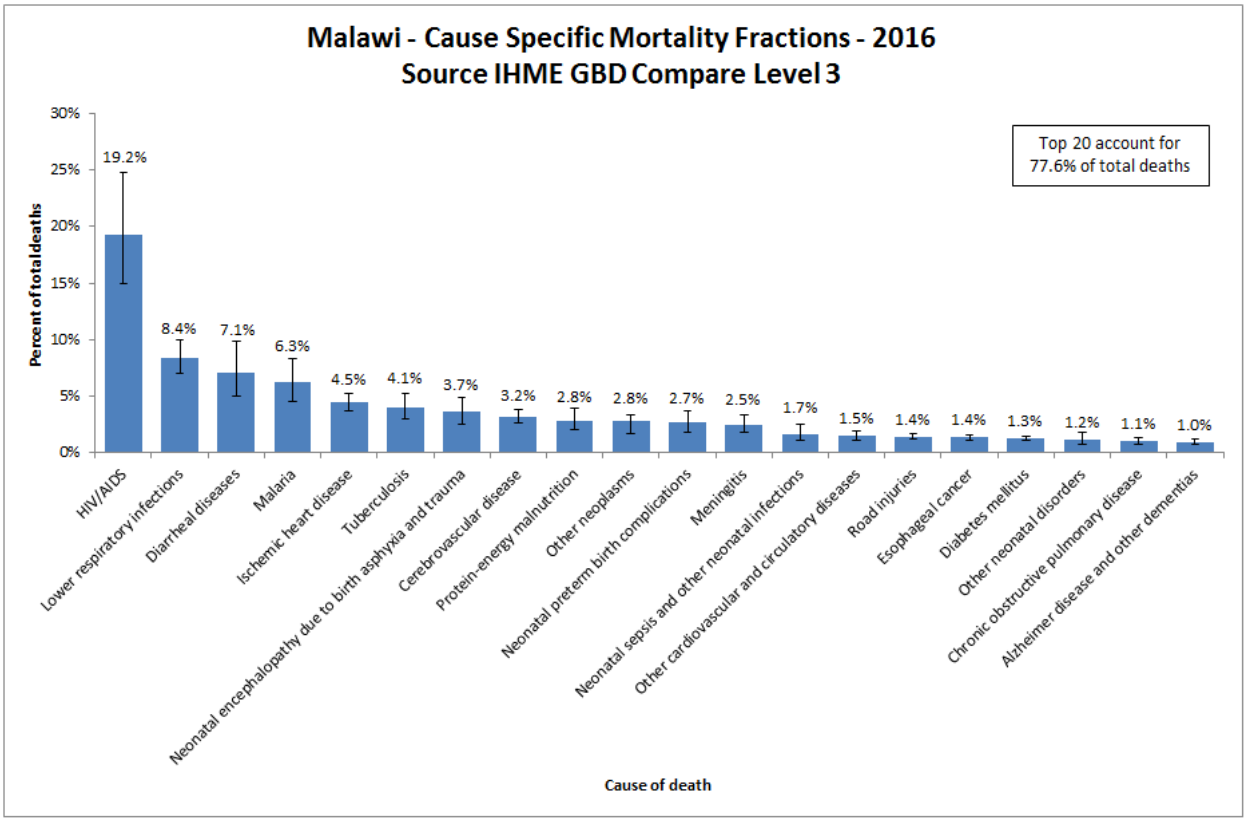

Colombia- Cause Specific Mortality Fractions - 2016
Source IHME GBD Compare Level 3
Top 20 account for 73.1% of total deaths
Percent of total deaths
30%
25%
20%
15%
10%
5%
0%
17.8%
7.6%
6.0%
5.6%
5.4%
3.6%
3.1%
3.1%
2.7%
2.5%
2.4%
2.2%
1.7%
1.5%
1.5%
1.4%
1.3%
1.3%
1.2%
1.1%
Ischemic heart disease
Interpersonal violence
Chronic obstructive pulmonary disease
Alzheimer disease and other dementias
Cerebrovascular disease
Chronic kidney disease
Diabetes mellitus
Road injuries
Lower respiratory infections
Hypertensive heart disease
Stomach cancer
Tracheal, bronchus, and lung cancer
Colon and rectum cancer
Prostate cancer
Other cardiovascular and circulatory diseases
Breast cancer
Other neoplasms
Self-harm
Congenital birth defects
HIV/AIDS
Cause of death

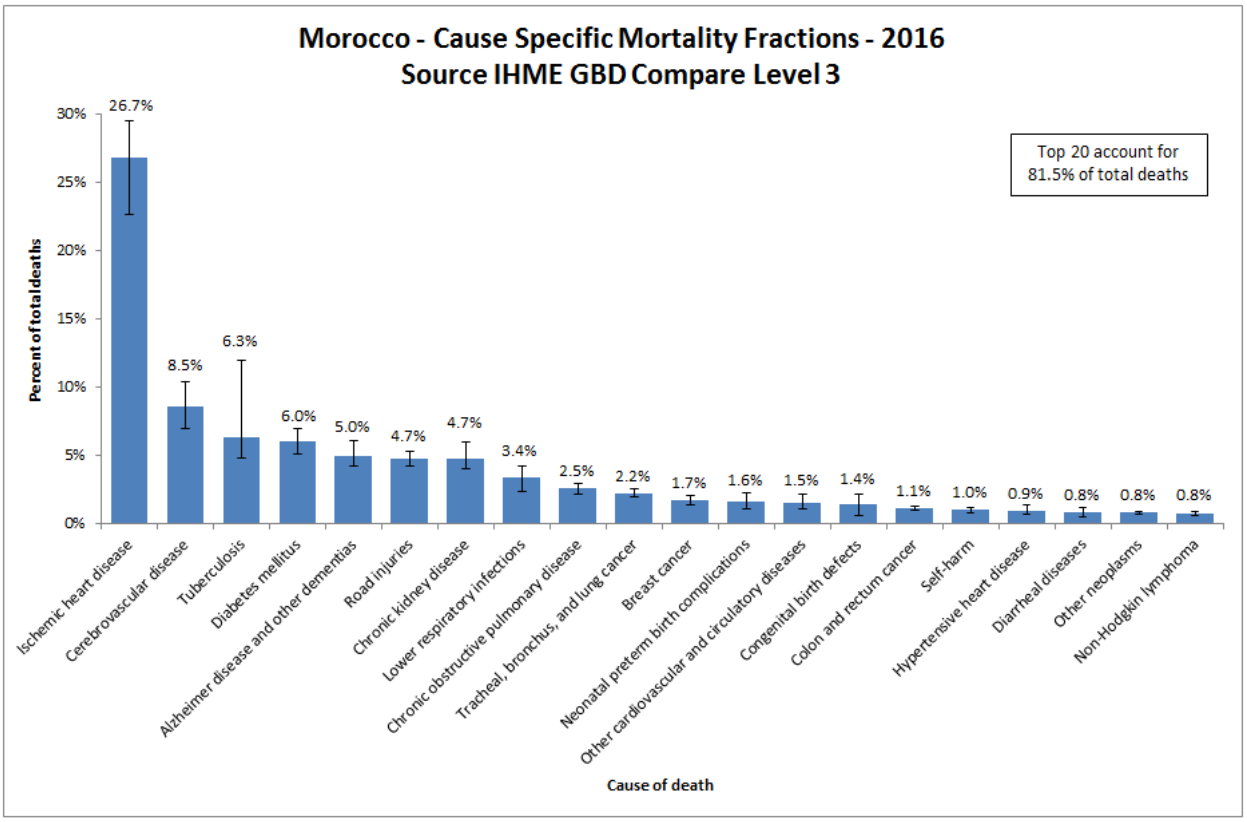

Ecuador - Cause Specific Mortality Fractions - 2016
Source IHME GBD Compare Level 3
Top 20 account for 69.3% of total deaths
Percent of total deaths
30%
25%
20%
15%
10%
5%
0%
11.3%
6.4%
5.9%
5.4%
5.4%
5.1%
4.0%
3.0%
3.0%
2.7%
2.7%
2.0%
1.8%
1.8%
1.7%
1.7%
1.6%
1.4%
1.3%
1.3%
Ischemic heart disease
Chronic kidney disease
Lower respiratory infections
Diabetes mellitus
Cerebrovascular disease
Road injuries
Alzheimer disease and other dementias
Stomach cancer
Interpersonal violence
Chronic obstructive pulmonary disease
Hypertensive heart disease
Neonatal preterm birth complications
Exposure to forces of nature
Congenital birth defects
Cirrhosis and other chronic liver diseases due to alcohol use
Self-harm
Prostate cancer
Tracheal, bronchus, and lung cancer
HIV/AIDS
Liver cancer
Cause of death

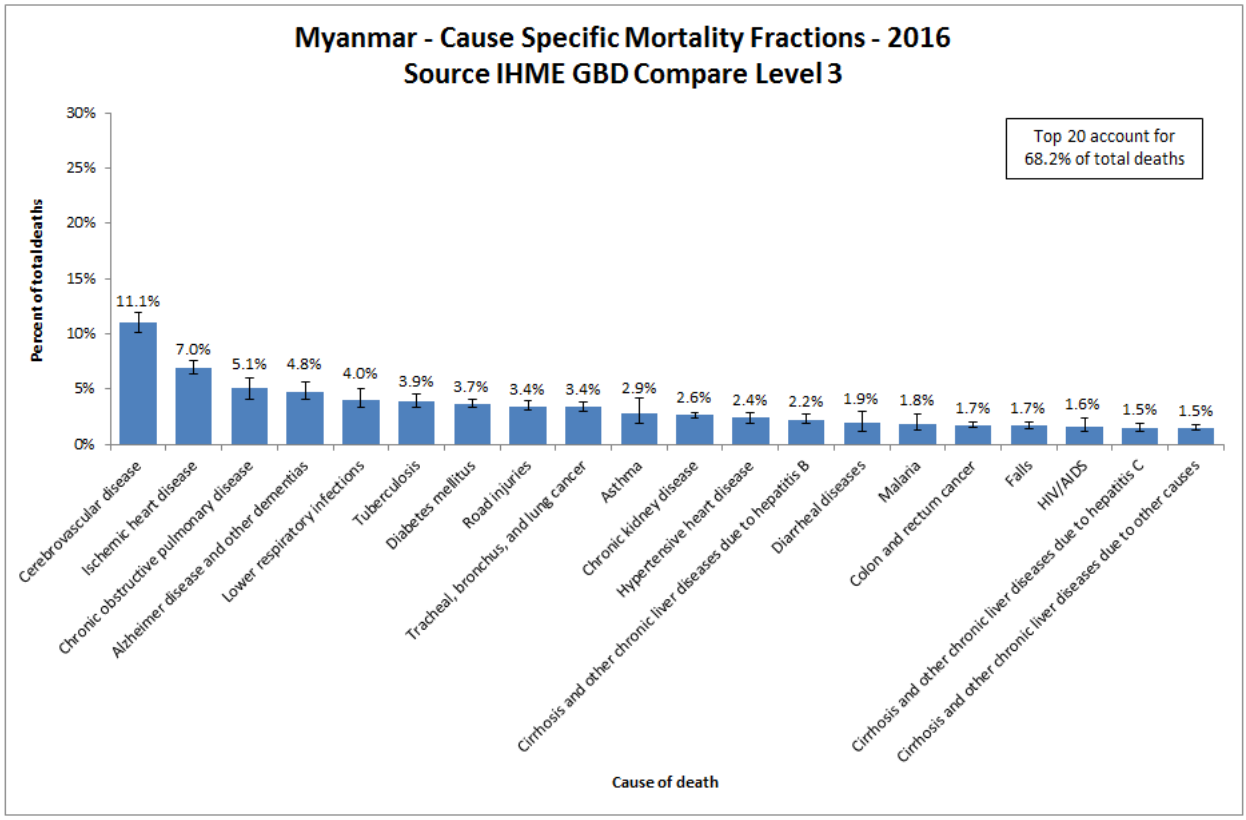

Ghana - Cause Specific Mortality Fractions - 2016
Source IHME GBD Compare Level 3
Top 20 account for
72.8% of total deaths
Percentage of total deaths
30%
25%
20%
15%
10%
5%
0%
10.4%
8.8%
7.7%
7.4%
7.4%
4.2%
3.3%
2.5%
2.5%
2.4%
2.3%
2.2%
1.9%
1.9%
1.5%
1.4%
1.4%
1.3%
1.3%
1.1%
Malaria
Lower respiratory infections
Ischemic heart disease
HIV/AIDS
Cerebrovascular disease
Neonatal encephalopathy due to birth asphyxia and trauma
Tuberculosis
Meningitis
Diabetes mellitus
Congenital birth defects
Diarrheal diseases
Neonatal sepsis and other neonatal infections
Neonatal preterm birth complications
Road injuries
Cirrhosis and other chronic liver diseases due to hepatitis B
Protein-energy malnutrition
Liver cancer
Chronic obstructive pulmonary disease
Chronic kidney disease
Other cardiovascular and circulatory diseases
Cause of death

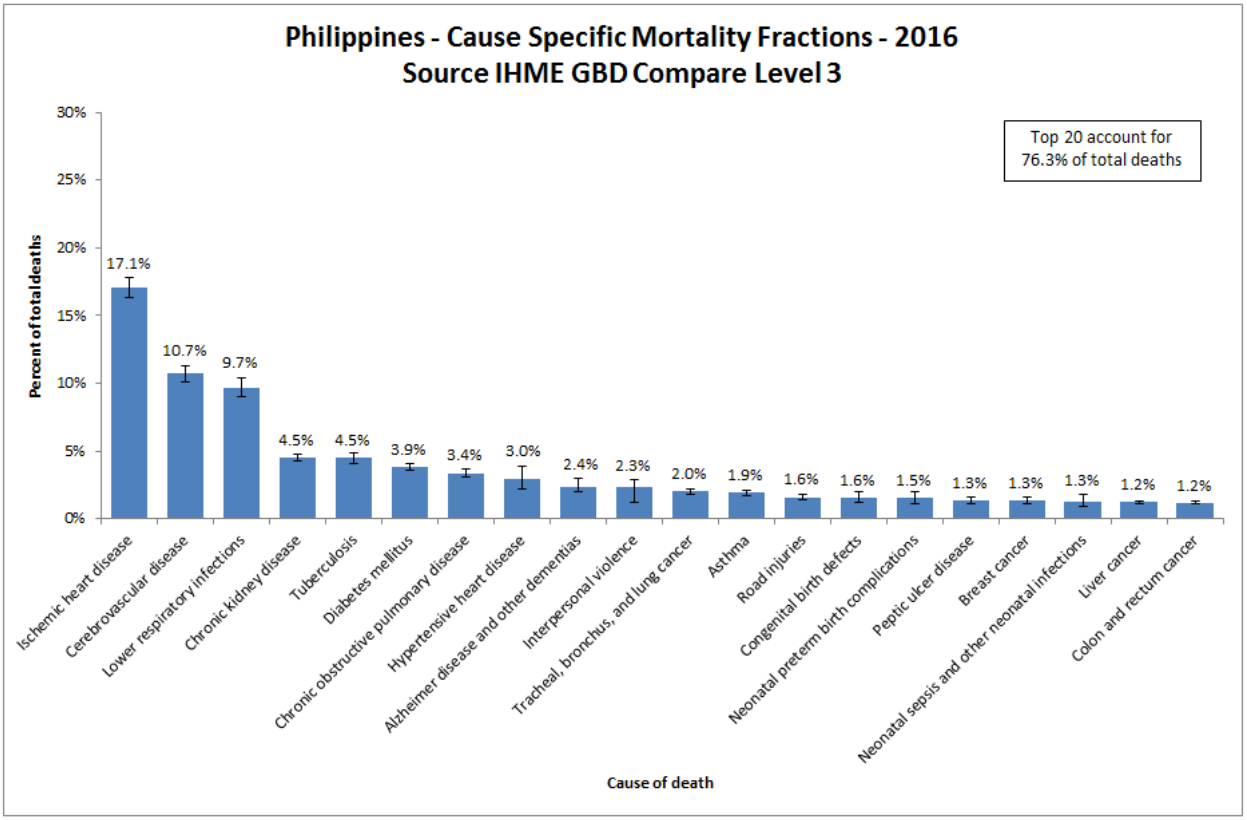

Indonesia - Cause Specific Mortality Fractions - 2016
Source IHME GBD Compare Level 3
Top 20 account for
73.4% of total deaths
Percent of total deaths
30%
25%
20%
15%
10%
5%
0%
17.8%
13.8%
5.8%
5.8%
3.2%
3.2%
3.0%
2.7%
2.5%
2.3%
2.1%
2.0%
1.5%
1.4%
1.1%
1.1%
1.1%
1.0%
1.0%
1.0%
Ischemic heart disease
Cerebrovascular disease
Tuberculosis
Diabetes mellitus
Chronic obstructive pulmonary disease
Diarrheal diseases
Alzheimer disease and other dementias
Road injuries
Lower respiratory infections
Chronic kidney disease
Asthma
Hypertensive heart disease
Neonatal preterm birth complications
Peptic ulcer disease
Neonatal encephalopathy due to birth asphyxia and trauma
Breast cancer
Falls
Cirrhosis and other chronic liver diseases due to other causes
Paralytic ileus and intestinal obstruction
Congenital birth defects
Cause of death

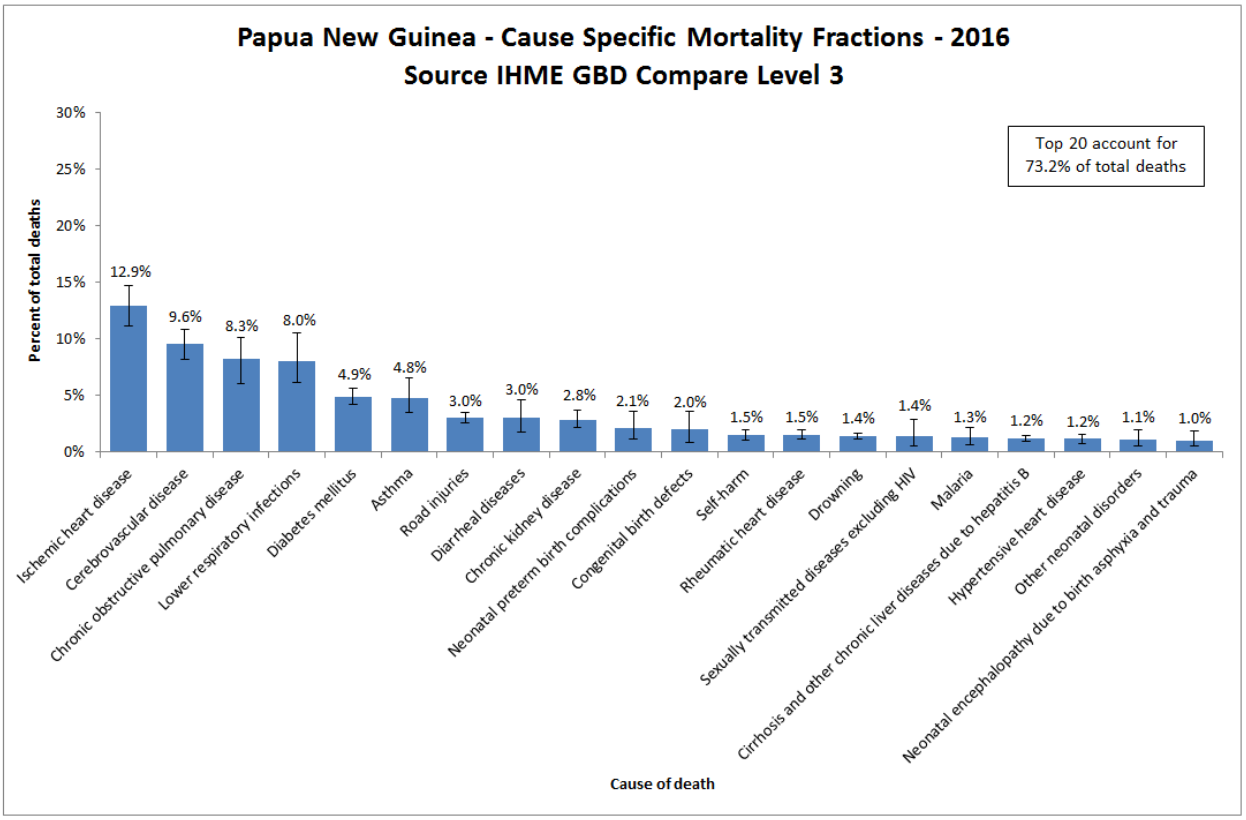

Kenya - Cause Specific Mortality Fractions - 2016
Source IHME GBD Compare Level 3
Top 20 account for 76.5% of total deaths
Percent of total deaths
30%
25%
20%
15%
10%
5%
0%
18.5%
15.6%
8.6%
4.0%
3.9%
3.7%
2.5%
2.1%
2.0%
2.0%
1.7%
1.7%
1.5%
1.4%
1.4%
1.3%
1.3%
1.2%
1.2%
1.0%
Diarrheal diseases
HIV/AIDS
Lower respiratory infections
Ischemic heart disease
Cerebrovascular disease
Tuberculosis
Neonatal encephalopathy due to birth asphyxia and trauma
Neonatal preterm birth complications
Meningitis
Protein-energy malnutrition
Malaria
Other cardiovascular and circulatory diseases
Road injuries
Tetanus
Interpersonal violence
Neonatal sepsis and other neonatal infections
Diabetes mellitus
Chronic obstructive pulmonary disease
Alzheimer disease and other dementias
Paralytic ileus and intestinal obstruction
Cause of death

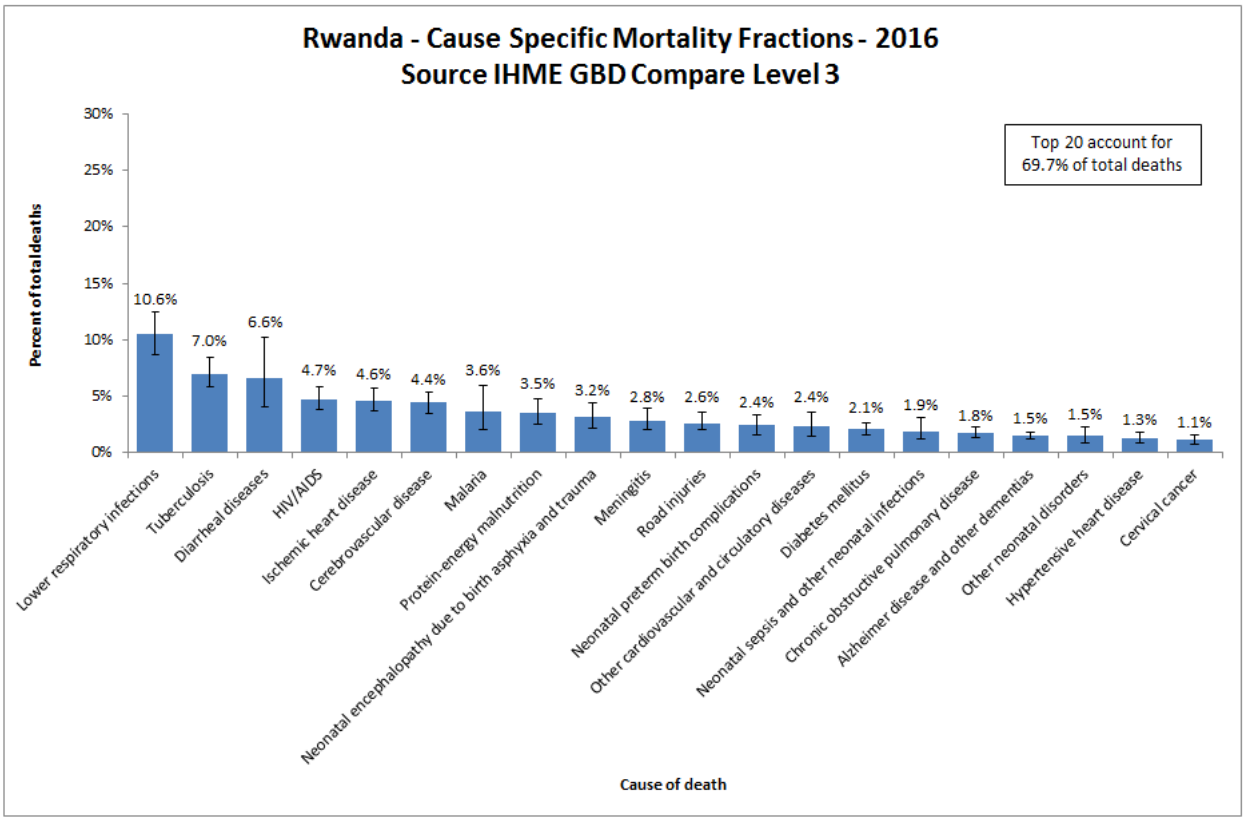

Malawi - Cause Specific Mortality Fractions - 2016
Source IHME GBD Compare Level 3
Top 20 account for 77.6% of total deaths
Percent of total deaths
30%
25%
20%
15%
10%
5%
0%
19.2%
8.4%
7.1%
6.3%
4.5%
4.1%
3.7%
3.2%
2.8%
2.8%
2.7%
2.5%
1.7%
1.5%
1.4%
1.4%
1.3%
1.2%
1.1%
1.0%
HIV/AIDS
Lower respiratory infections
Diarrheal diseases
Malaria
Ischemic heart disease
Tuberculosis
Neonatal encephalopathy due to birth asphyxia and trauma
Cerebrovascular disease
Protein-energy malnutrition
Other neoplasms
Neonatal preterm birth complications
Meningitis
Neonatal sepsis and other neonatal infections
Other cardiovascular and circulatory diseases
Road injuries
Esophageal cancer
Diabetes mellitus
Other neonatal disorders
Chronic obstructive pulmonary disease
Alzheimer disease and other dementias
Cause of death

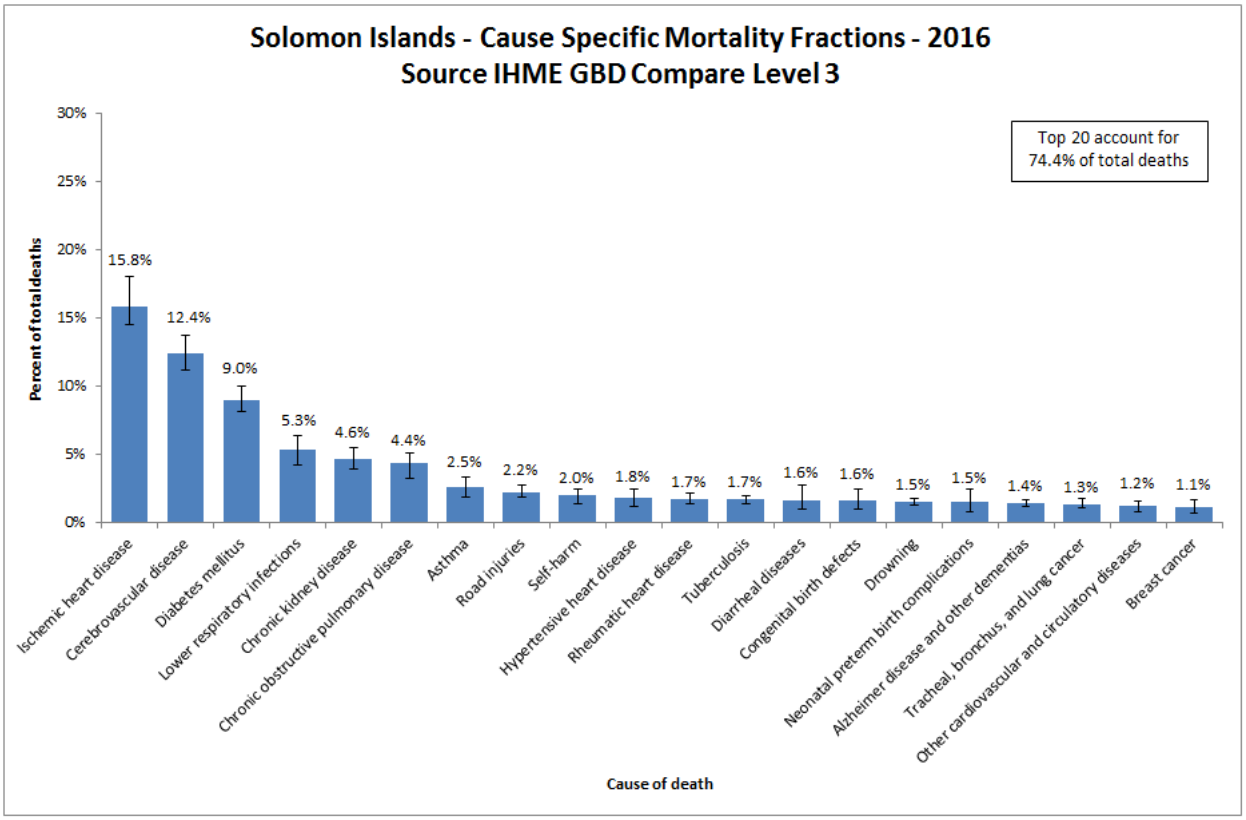

Morocco - Cause Specific Mortality Fractions - 2016
Source IHME GBD Compare Level 3
Top 20 account for 81.5% of total deaths
Percent of total deaths
30%
25%
20%
15%
10%
5%
0%
26.7%
8.5%
6.3%
6.0%
5.0%
4.7%
4.7%
3.4%
2.5%
2.2%
1.7%
1.6%
1.5%
1.4%
1.1%
1.0%
0.9%
0.8%
0.8%
0.8%
Ischemic heart disease
Cerebrovascular disease
Tuberculosis
Diabetes mellitus
Alzheimer disease and other dementias
Road injuries
Chronic kidney disease
Lower respiratory infections
Chronic obstructive pulmonary disease
Tracheal, bronchus, and lung cancer
Breast cancer
Neonatal preterm birth complications
Other cardiovascular and circulatory diseases
Congenital birth defects
Colon and rectum cancer
Self-harm
Hypertensive heart disease
Diarrheal diseases
Other neoplasms
Non-Hodgkin lymphoma
Cause of death

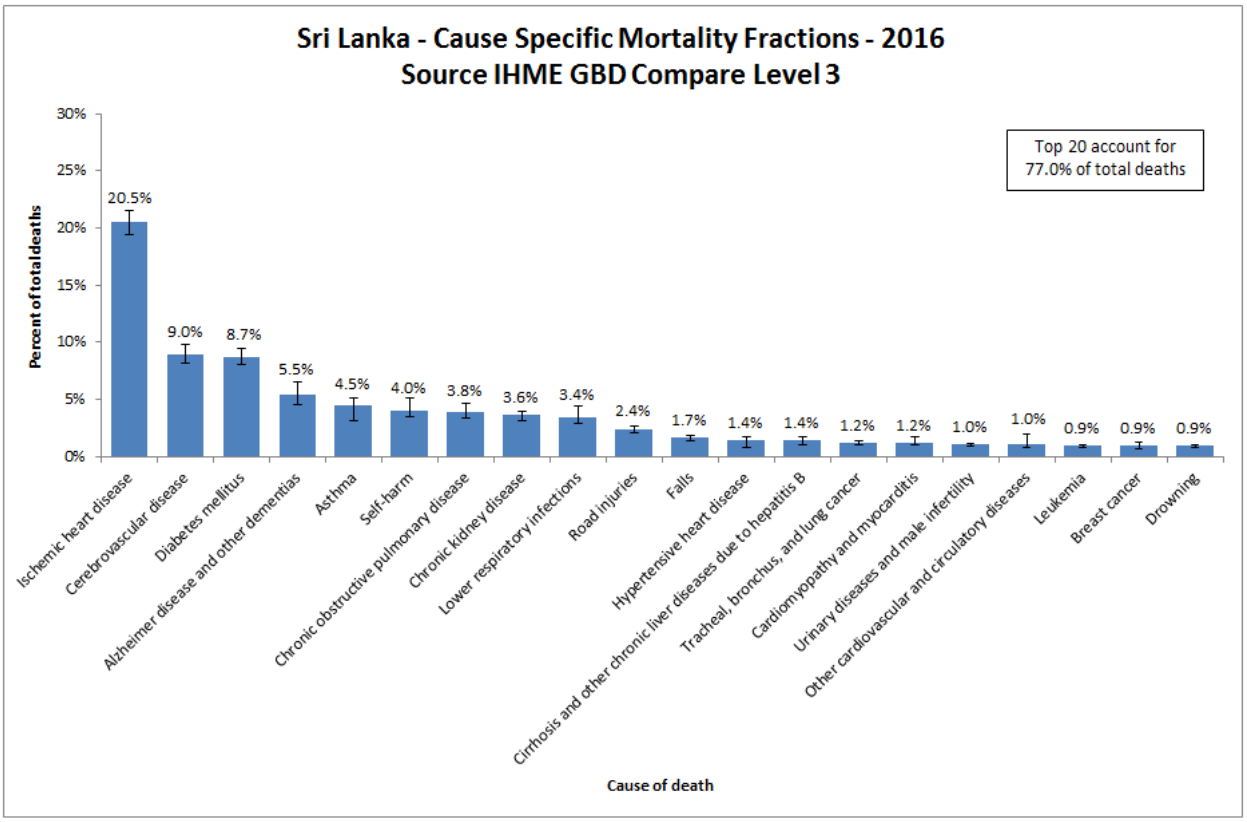

Myanmar - Cause Specific Mortality Fractions - 2016
Source IHME GBD Compare Level 3
Top 20 account for 68.2% of total deaths
Percent of total deaths
30%
25%
20%
15%
10%
5%
0%
11.1%
7.0%
5.1%
4.8%
4.0%
3.9%
3.7%
3.4%
3.4%
2.9%
2.6%
2.4%
2.2%
1.9%
1.8%
1.7%
1.7%
1.6%
1.5%
1.5%
Cerebrovascular disease
Ischemic heart disease
Chronic obstructive pulmonary disease
Alzheimer disease and other dementias
Lower respiratory infections
Tuberculosis
Diabetes mellitus
Road injuries
Tracheal, bronchus, and lung cancer
Asthma
Chronic kidney disease
Hypertensive heart disease
Cirrhosis and other chronic liver diseases due to hepatitis B
Diarrheal diseases
Malaria
Colon and rectum cancer
Falls
HIV/AIDS
Cirrhosis and other chronic liver diseases due to hepatitis C
Cirrhosis and other chronic liver diseases due to other causes
Cause of death

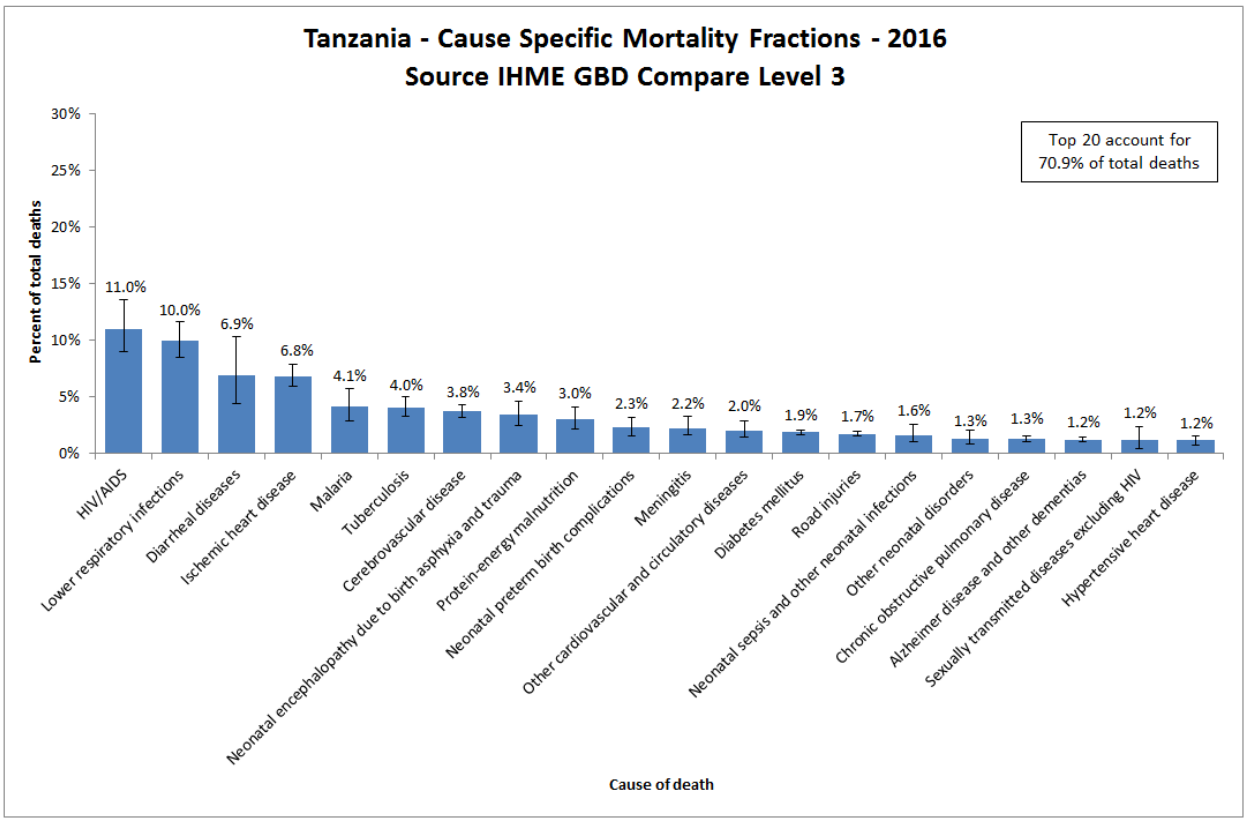

Philippines - Cause Specific Mortality Fractions - 2016
Source IHME GBD Compare Level 3
Top 20 account for 76.3% of total deaths
Percent of total deaths
30%
25%
20%
15%
10%
5%
0%
17.1%
10.7%
9.7%
4.5%
4.5%
3.9%
3.4%
3.0%
2.4%
2.3%
2.0%
1.9%
1.6%
1.6%
1.5%
1.3%
1.3%
1.3%
1.2%
1.2%
Ischemic heart disease
Cerebrovascular disease
Lower respiratory infections
Chronic kidney disease
Tuberculosis
Diabetes mellitus
Chronic obstructive pulmonary disease
Hypertensive heart disease
Alzheimer disease and other dementias
Interpersonal violence
Tracheal, bronchus, and lung cancer
Asthma
Road injuries
Congenital birth defects
Neonatal preterm birth complications
Peptic ulcer disease
Breast cancer
Neonatal sepsis and other neonatal infections
Liver cancer
Colon and rectum cancer
Cause of death

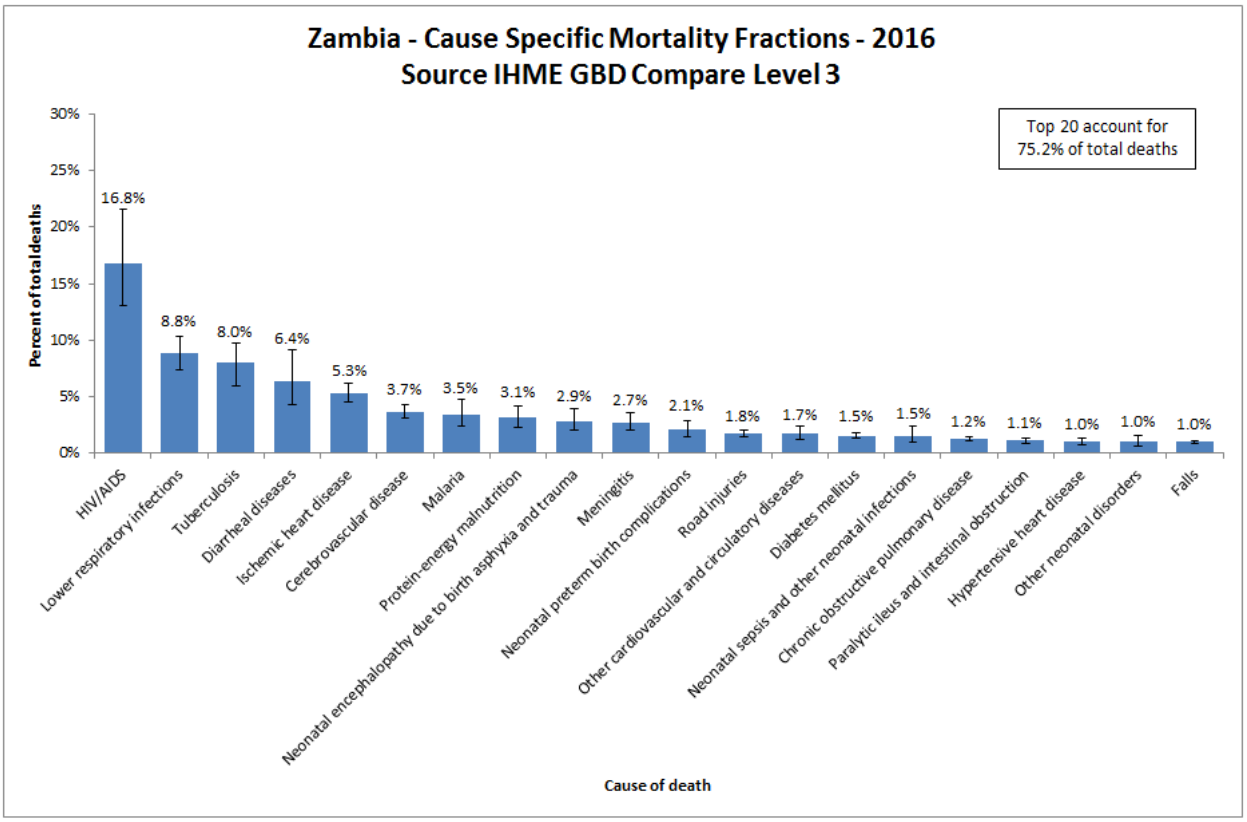

Papua New Guinea - Cause Specific Mortality Fractions - 2016
Source IHME GBD Compare Level 3
Top 20 account for 73.2% of total deaths
Percent of total deaths
30%
25%
20%
15%
10%
5%
0%
12.9%
9.6%
8.3%
8.0%
4.9%
4.8%
3.0%
3.0%
2.8%
2.1%
2.0%
1.5%
1.5%
1.4%
1.4%
1.3%
1.2%
1.2%
1.1%
1.0%
Ischemic heart disease
Cerebrovascular disease
Chronic obstructive pulmonary disease
Lower respiratory infections
Diabetes mellitus
Asthma
Road injuries
Diarrheal diseases
Chronic kidney disease
Neonatal preterm birth complications
Congenital birth defects
Self-harm
Rheumatic heart disease
Drowning
Sexually transmitted diseases excluding HIV
Malaria
Cirrhosis and other chronic liver diseases due to hepatitis B
Hypertensive heart disease
Other neonatal disorders
Neonatal encephalopathy due to birth asphyxia and trauma
Cause of death


Rwanda - Cause Specific Mortality Fractions - 2016
Source IHME GBD Compare Level 3
Top 20 account for
69.7% of total deaths
Percent of total deaths
30%
25%
20%
15%
10%
5%
0%
10.6%
7.0%
6.6%
4.7%
4.6%
4.4%
3.6%
3.5%
3.2%
2.8%
2.6%
2.4%
2.4%
2.1%
1.9%
1.8%
1.5%
1.5%
1.3%
1.1%
Lower respiratory infections
Tuberculosis
Diarrheal diseases
HIV/AIDS
Ischemic heart disease
Cerebrovascular disease
Malaria
Protein-energy malnutrition
Neonatal encephalopathy due to birth asphyxia and trauma
Meningitis
Road injuries
Neonatal preterm birth complications
Other cardiovascular and circulatory diseases
Diabetes mellitus
Neonatal sepsis and other neonatal infections
Chronic obstructive pulmonary disease
Alzheimer disease and other dementias
Other neonatal disorders
Hypertensive heart disease
Cervical cancer
Cause of death


Solomon Islands - Cause Specific Mortality Fractions - 2016
Source IHME GBD Compare Level 3
Top 20 account for
74.4% of total deaths
Percent of total deaths
30%
25%
20%
15%
10%
5%
0%
15.8%
12.4%
9.0%
5.3%
4.6%
4.4%
2.5%
2.2%
2.0%
1.8%
1.7%
1.7%
1.6%
1.6%
1.5%
1.5%
1.4%
1.3%
1.2%
1.1%
Ischemic heart disease
Cerebrovascular disease
Diabetes mellitus
Lower respiratory infections
Chronic kidney disease
Chronic obstructive pulmonary disease
Asthma
Road injuries
Self-harm
Hypertensive heart disease
Rheumatic heart disease
Tuberculosis
Diarrheal diseases
Congenital birth defects
Drowning
Neonatal preterm birth complications
Alzheimer disease and other dementias
Tracheal, bronchus, and lung cancer
Other cardiovascular and circulatory diseases
Breast cancer
Cause of death


Sri Lanka - Cause Specific Mortality Fractions - 2016
Source IHME GBD Compare Level 3
Top 20 account for
77.0% of total deaths
Percent of total deaths
30%
25%
20%
15%
10%
5%
0%
20.5%
9.0%
8.7%
5.5%
4.5%
4.0%
3.8%
3.6%
3.4%
2.4%
1.7%
1.4%
1.4%
1.2%
1.2%
1.0%
1.0%
0.9%
0.9%
0.9%
Ischemic heart disease
Cerebrovascular disease
Diabetes mellitus
Alzheimer disease and other dementias
Asthma
Self-harm
Chronic obstructive pulmonary disease
Chronic kidney disease
Lower respiratory infections
Road injuries
Falls
Hypertensive heart disease
Cirrhosis and other chronic liver diseases due to hepatitis B
Tracheal, bronchus, and lung cancer
Cardiomyopathy and myocarditis
Urinary diseases and male infertility
Other cardiovascular and circulatory diseases
Leukemia
Breast cancer
Drowning
Cause of death


Tanzania - Cause Specific Mortality Fractions - 2016
Source IHME GBD Compare Level 3
Top 20 account for
70.9% of total deaths
Percent of total deaths
30%
25%
20%
15%
10%
5%
0%
11.0%
10.0%
6.9%
6.8%
4.1%
4.0%
3.8%
3.4%
3.0%
2.3%
2.2%
2.0%
1.9%
1.7%
1.6%
1.3%
1.3%
1.2%
1.2%
1.2%
HIV/AIDS
Lower respiratory infections
Diarrheal diseases
Ischemic heart disease
Malaria
Tuberculosis
Cerebrovascular disease
Neonatal encephalopathy due to birth asphyxia and trauma
Protein-energy malnutrition
Neonatal preterm birth complications
Meningitis
Other cardiovascular and circulatory diseases
Diabetes mellitus
Road injuries
Neonatal sepsis and other neonatal infections
Other neonatal disorders
Chronic obstructive pulmonary disease
Alzheimer disease and other dementias
Sexually transmitted diseases excluding HIV
Hypertensive heart disease
Cause of death


Zambia - Cause Specific Mortality Fractions - 2016
Source IHME GBD Compare Level 3
Top 20 account for 75.2% of total deaths
Percent of total deaths
30%
25%
20%
15%
10%
5%
0%
16.8%
8.8%
8.0%
6.4%
5.3%
3.7%
3.5%
3.1%
2.9%
2.7%
2.1%
1.8%
1.7%
1.5%
1.5%
1.2%
1.1%
1.0%
1.0%
1.0%
HIV/AIDS
Lower respiratory infections
Tuberculosis
Diarrheal diseases
Ischemic heart disease
Cerebrovascular disease
Malaria
Protein-energy malnutrition
Neonatal encephalopathy due to birth asphyxia and trauma
Meningitis
Neonatal preterm birth complications
Road injuries
Other cardiovascular and circulatory diseases
Diabetes mellitus
Neonatal sepsis and other neonatal infections
Chronic obstructive pulmonary disease
Paralytic ileus and intestinal obstruction
Hypertensive heart disease
Other neonatal disorders
Falls
Cause of death

C.4. Verbal Autopsy Target Cause Lists

	WHO 2016 VA (64 causes & 70 codes)	ICD-10	IHME SmartVA (47 causes, 46 codes)	ICD-10
1	Diarrheal diseases	A09	Diarrhea/Dysentery	A09
2	Pulmonary tuberculosis	A16	TB	A16
3	Neonatal tetanus	A33	Sepsis	A41
4	Tetanus	A34	Hemorrhagic fever	A99
5	Tetanus Obstetric	A35	Measles	B05
6	Pertussis	A37	AIDS	B24
7	Sepsis	A41	Malaria	B54
8	Dengue fever	A90	Other Infectious Diseases	B99
9	Haemorrhagic fever	A99	Cancer Esophageal	C15
10	Measles	B05	Cancer Stomach	C16
11	HIV/AIDS related death	B24	Cancer colorectal	C18
12	Malaria	B54	Cancer Lung	C34
13	Unspecified infectious disease	B99	Cancer Breast	C50
14	Oral neoplasm	C06	Cancer Cervical	C53
15	Digestive neoplasms	C26	Cancer Prostrate	C61
16	Respiratory neoplasms	C39	Cancers Other	C76
17	Breast neoplasms	C50	Leukemia/Lymphomas	C96
18	Female reproductive neoplasms	C57	Diabetes	E14
19	Male reproductive neoplasms	C63	Meningitis	G03
20	Other and unspecified neoplasms	C80	Encephalitis	G04
21	Sickle cell with crisis	D57	Epilepsy	G40
22	Severe anemia	D64	Stroke	I64
23	Diabetes mellitus	E14	Other Cardiovascular Diseases	I99
24	Severe malnutrition	E46	Pneumonia	J22
25	Meningitis and encephalitis	G03	Pneumonia (newborn)	P23
26	Meningitis and encephalitis	G04	Cirrhosis	K74
27	Epilepsy	G40	Other Digestive Diseases	K92
28	Acute cardiac disease (ischemic)	I24	Renal Failure	N19
29	Stroke	I64	Maternal	O95
30	Other and unspecified cardiac disease	I99	Preterm Delivery	P07
31	Acute respiratory infection, Pneumonia	J18	Birth asphyxia	P21
32	Acute respiratory infection, Pneumonia	J22	Meningitis/Sepsis	P36
33	Chronic obstructive pulmonary disease (COPD)	J44	Stillbirth	P95
34	Asthma	J45	Congenital malformation	Q89
35	Liver cirrhosis	K74	Road Traffic	V89
36	Renal failure	N19	Falls	W19
37	Ectopic pregnancy	O00	Drowning	W74
38	Other and unspecified maternal cause	O05	Fires	X09
39	Abortion-related death	O06	Bite of Venomous Animal	X27
40	Pregnancy-induced hypertension	O13	Poisonings	X49
41	Pregnancy-induced hypertension (eclampsia)	O15	Other Injuries	X58
42	Obstetric haemorrhage (ante partum)	O46	Suicide	X84
43	Obstetric labour	O66	Violent Death / Homicide	Y09
44	Ruptured uterus	O71	Chronic Respiratory	J44
45	Obstetric haemorrhage (post partum)	O72	Ischemic Heart Disease	I24
46	Pregnancy-related sepsis (ante partum)	O75	Other Defined Causes of Child Deaths	R99
47	Pregnancy-related sepsis (post partum)	O85	Other Non-communicable Diseases	R99
48	Anemia of pregnancy	O99		
49	Prematurity	P07		
50	Birth asphyxia	P21		
51	Neonatal pneumonia	P23		
52	Neonatal sepsis	P63		
53	Fresh stillbirth	P95		
54	Macerated stillbirth	P95		
55	Other and unspecified perinatal cause of death	P96		
56	Congenital malformation	Q89		
57	Acute abdomen	R10		
59	Other and unspecified non-communicable disease	R99		
58	Cause of death unknown	R99		
60	Road traffic accident	V89		
61	Other transport accident	V99		
62	Accidental fall	W19		
63	Accidental drowning and submersion	W74		
64	Accidental exposure to smoke, fire and flames	X09		
65	Contact with venomous animals and plants	X29		
66	Exposure to force of nature	X39		
67	Accidental poisoning and exposure to noxious substa	X49		
68	Other and unspecified external cause of death	X59		
69	Intentional self-harm	X84		
70	Assault	Y09		

C.5. Link to CRVS VA Costing Tool for download

To download the CRVS VA Costing tool, please follow the subsequent link:

https://crvsgateway.info/learningcentre/improving-quality-and-presentation-of-crvs-data/verbal-autopsy-costing-and-budgeting-tool

References

Bierrenbach, A. (2008). Steps in applying Probability Proportional to Size (PPS) and calculating Basic Probability Weights http://www.who.int/tb/advisory_bodies/impact_measurement_taskforce/meetings/prevalence_survey/psws_probability_prop_size_bierrenbach.pdf, World Health Organization, Geneva, Switzerland

D'Ambruoso, L., T. Boerma, P. Byass, E. Fottrell, K. Herbst, K. Kallander and Z. Mullan (2016). "The case for verbal autopsy in health systems strengthening." Lancet Glob Health.

de Savigny, D., S. Renggli, D. Cobos Muñoz and M. Collinson (2017). Maximizing Synergies between Health Observatories and CRVS: Guidance for INDEPTH HDSS Sites and Other CRVS Stakeholders. INDEPTH Network and Bloomberg Philanthropies Data for Health Initiative. https://crvsgateway.info/.

de Savigny, D., I. Riley, D. Chandramohan, F. Odhiambo, E. Nichols, S. Notzon, C. AbouZahr, R. Mitra, D. Cobos Munoz, S. Firth, N. Maire, O. Sankoh, G. Bronson, P. Setel, P. Byass, R. Jakob, T. Boerma and A. D. Lopez (2017). "Integrating community-based verbal autopsy into civil registration and vital statistics (CRVS): system-level considerations." Glob Health Action **10**(1): 1272882.

Eldridge, S. M., D. Ashby and S. Kerry (2006). "Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method." Int J Epidemiol **35**(5): 1292-1300.

Everitt, B. S. and A. Skrondal (2010). The Cambridge Dictionary of Statistics. New York, Cambridge University Press. **4th**.

Hayes, J. and L. Moulton (2009). Cluster Randomized Trails. Boca Raton, CRC Press.

Hayes, R. J., N. D. Alexander, S. Bennett and S. N. Cousens (2000). "Design and analysis issues in cluster-randomized trials of interventions against infectious diseases." Stat Methods Med Res **9**(2): 95-116.

Hayes, R. J. and S. Bennett (1999). "Simple sample size calculation for cluster-randomized trials." Int J Epidemiol **28**(2): 319-326.

International Epidemiological Association (2014). A Dictionary of Epidemiology. M. Porta, S. Greenland, M. Hernán, I. dos Santos Silva and J. M. Last. New York, Oxford University Press. **6th**.

Kerry, S. M. and J. M. Bland (1998). "The intracluster correlation coefficient in cluster randomisation." BMJ **316**(7142): 1455.

Killip, S., Z. Mahfoud and K. Pearce (2004). "What is an intracluster correlation coefficient? Crucial concepts for primary care researchers." Ann Fam Med **2**(3): 204-208.

National Bureau of Statistics and Ministry of Finance (2013). 2012 Population and housing census. Dar es Salaam, National Bureau of Statistics and Ministry of Finance, United Republic of Tanzania.

Pagel, C., A. Prost, S. Lewycka, S. Das, T. Colbourn, R. Mahapatra, K. Azad, A. Costello and D. Osrin (2011). "Intracluster correlation coefficients and coefficients of variation for perinatal outcomes from five cluster-randomised controlled trials in low and middle-income countries: results and methodological implications." Trials **12**: 151.

Preisser, J. S., B. A. Reboussin, E. Y. Song and M. Wolfson (2007). "The importance and role of intracluster correlations in planning cluster trials." Epidemiology **18**(5): 552-560.

Upton, G. and I. Cook (2014). A Dictionary of Statistics. Oxford, Oxford University Press. **3rd**.