DOCUMENTS DE TRAVAIL 268

Studying multiple causes of death in LMICs in the absence of death certificates : taking advantage of probabilistic cause-of-death estimation methods (InterVA-4)

Ariane Sessego

Directed by Géraldine Duthé,
with the collaboration of Bruno Lankoandé,
Dianou Kassoum and Serge Rabier

Septembre 2021

ined
INSTITUT NATIONAL
D'ÉTUDES DÉMOGRAPHIQUES

Ariane Sessego, directed by Géraldine Duthé, with the collaboration of Bruno Lankoandé, Dianou Kassoum and Serge Rabier, *Studying multiple causes of death in LMICs in the absence of death certificates : taking advantage of probabilistic cause-of-death estimation methods (InterVA-4)*, Paris, Ined, Document de travail, 268

Studying multiple causes of death in LMICs in the absence of death certificates : taking advantage of probabilistic cause-of-death estimation methods (InterVA-4)

Internship report - September 2020 to May 2021 - INED

Directed by Géraldine DUTHE (INED[1])
With the collaboration of Bruno Lankoandé and Dianou Kassoum (ISSP[2]) and Serge Rabier[3]
Funded by AFD[4]

Ariane Sessego

[1]Institut National d'Etudes Démographiques, Paris
[2]Institut Supérieur des Sciences des Populations, Ouagadougou, Burkina Faso
[3]Agence Française du Développement, Paris
[4]Agence Française du Développement, Paris

.

Abstract

In low- and middle-income countries (LMICs), the burden of non-communicable diseases is increasing due to the combination of population aging and lifestyle changes. While interest in multimorbidity has been rising to study more precisely the complex morbid processes that adults experience, health data in LMICs are scarce and rarely allow such investigations. Focusing on multimorbidity leading to death, we aim to develop an approach to estimate multiple causes of death using available data. In settings where certification of death by physicians is not available, verbal autopsies (VAs) have been developed to diagnose likely causes of death from information collected via a structured interview with final caregivers about the signs and symptoms leading up to death. With an increasing use of probabilistic models to interpret VAs, we investigate their potential for identifying multiple causes using a database of 72,330 adult deaths (15 and older) from 22 Health and Demographic Surveillance System (HDSS) sites located in Asia and Africa, and detailed VA data from the Ouagadougou HDSS in Burkina Faso (1,700 deaths). The Bayesian model InterVA-4 attributes multiple likely causes to 11% of deaths. However, some combinations result more from uncertain diagnosis than from multimorbidity. Elaborating an index of similarity between causes based on the InterVA's probability matrix, we aim to differentiate competing causes (uncertainty) from co-occurring causes (multimorbidity). Selecting the most dissimilar associations of causes, we highlight the importance of associations between infectious and non-communicable diseases, as well as the burden of diabetes and cardiovascular diseases among the identified multimorbidity.

Key words: multimorbidity, cause-of-death statistics, low and middle income countries, data quality, verbal autopsies, HDSS

Résumé

Dans les pays à revenu faible et intermédiaire, le poids des maladies non transmissibles et chroniques augmente sous l'effet conjugué du vieillissement et des changements de mode de vie. Chez les adultes particulièrement, les processus morbides sont complexes. Si l'intérêt pour la multimorbidité s'est fortement développé dans les pays vieillissants, les données sanitaires restent parcellaires dans les pays à revenu faible et intermédiaire, rendant son appréhension difficile. Nous intéressant particulièrement à la multimorbidité menant au décès, nous cherchons à développer une approche pour estimer les causes multiples de décès à partir des données disponibles. Dans les observatoires de population, quand un décès est enregistré, des autopsies verbales permettent de recueillir auprès des proches l'histoire de la maladie et des symptômes ayant conduit au décès afin d'en déterminer la cause probable. Le réseau INDEPTH met a disposition une base de décès par causes pour 22 sites, utilisant le logiciel InterVA-4 comme méthodologie unifiée pour déterminer les causes de décès. Sur les 72 330 décès de plus de 15 ans de la base, le modèle bayésien attribue plusieurs causes probables pour 11% des décès. Pour autant, certaines combinaisons sont plus liées à une incertitude de diagnostic qu'à de la multimorbidité. Par l'intermédiaire d'un indice de similarité établi sur la base des symptômes, nous cherchons à distinguer les causes co-occurrentes de celles qui se concurrencent. En sélectionnant les associations de causes les plus dissemblables, nous mettons en avant l'importance des associations entre maladies infectieuses et maladies non transmissibles, ainsi que le poids du diabète et des maladies cardiovasculaires dans la multimorbidité.

Mots clés : multimorbidité, causes multiples de décès, qualité des données, pays à revenu faible et intermédiaire, autopsies verbales, observatoires de population.

Contents

1. Background and main objectives

Low- and middle-income countries (LMICs) have been experiencing over the last 30 years to varying degrees an increase in life expectancy, marked by a decrease in the burden of infectious diseases responsible for high child mortality and an increase in the burden of non-communicable diseases (cardiovascular diseases, cancer, diabetes...). This so-called "epidemiological transition" underway attests to the progress of health care in LMICs, efforts to increase life expectancy that will hopefully carry on.

However, this transition, accompanied by the shift and the diversification of the burden of diseases, represents important challenges for LMICs that remain under-studied due to lack of data. Focusing on multimorbidity, this work aims to explore and present promising methods using cause-specific mortality data to analyse the health transition and inform health care systems, interpreted by the InterVA algorithm.

1.1 The health transition in LMICs

1.1.1 Definitions: epidemiological and health transition

First theorised by Abdel R. Omran in 1971 (Omran, 1971), **the epidemiological transition** refers to a transformation of the epidemiologic profile of populations, shifting from a sanitary situation dominated by infectious, neonatal, maternal and nutritional diseases to a profile where the main causes of death are considered "lifestyle diseases", mainly non-communicable diseases (e.g. cardiovascular diseases, cancers...) and injuries (most importantly road injuries).

Based on the experience of High Income Countries (HICs), Omran theorised three successive stages of this transition, closely linked to the demographic transition and the increase in life expectancy :

- During the first stage, **the "age of famine and pestilence"**, the mortality rates are high and fluctuating. Causes of death are dominated by infectious and nutritional diseases, with a life expectancy around 20 to 30 years old, and pandemics are frequent.

- The second stage, **the "age of receding pandemics"** marks the beginning of the epidemiological transition. It is characterised by a stabilisation of the mortality rate, and a significant decrease in neonatal mortality. Acute infectious diseases are no longer the main cause of mortality and life expectancy rises to about 45 to 55 years old.

- During the third stage, **the "age of degenerative and man-made diseases"**, mortality rates drop substantially, and life expectancy surpasses 70 years old. Infectious, nutritional, neonatal and maternal cause of death, also referred to as "diseases of poverty" give way to non-communicable diseases and violent causes of death as main causes of death, refered to as "man-made diseases" by Omran.

- A fourth stage could also be added : **the "age of delayed degenerative diseases"**, in order to take into account the considerable progress in cardiovascular mortality recorded during the 1970s in HICs (Olshanski and Ault, 1986), resulting from behavioural and medical improvement. During this stage, characteristic of higher income countries, most of the population manages to reach an advanced age, with older years characterised epidemiologically by degenerative diseases. Ageing and dependency become a societal challenge.

To this concept of epidemiological transition, focused on the evolution of of causes of death, we could substitute the broader notion of **"health transition"**, aiming to also capture the medical and behavioural evolutions accompanying these changes (Meslé and Vallin, 2002). The notion intends in addition to nuance Omran's linear model,

taking into account the diversity of trajectories, obstacles and possible drawbacks. This definition could allow us to identify two different historical processes that have been underlined as part of health transitions, which are not meant to be seen as part of a linear model, but rather as possibly overlapping shifts in medical and individual behaviours around health in different societies :

- One transition is characterised by the recession of "diseases of poverty" (infectious, nutritional, neonatal and maternal diseases), marked by medical advancements such as vaccination, neonatal and maternal care as well as improvements in sanitisation through individual behaviour (lead by the pasteurian and hygienist revolution in Europe and the USA at the turn of the 20th century, for example).

- A second transition would be the decline of cardiovascular diseases as observed in the 1970s in HICs, resulting from the rising awareness of risk factors leading to behavioural changes and development of new medical techniques.

Although the recent Covid-19 pandemic has been calling into question this linear model, it remains a reference for understanding the transformations of the epidemiological profile of populations. Moreover, elaborated from the epidemiological evolutions of HICs since the 17th century, it constitutes an interesting point of comparison and discussion for LMICs, where these evolutions have been considerably quicker.

1.1.2 The health transition and the different trajectories of LMICs

In LMICs, the health transition has been ongoing at least since the start of the twentieth century, but it is only around the 1950s that data starts to be recorded. It is accompanied with an increase in life expectancy, mainly due to the decrease in infant and child mortality. These transformations can be illustrated by the trends in life expectancy at birth from 1960 to 2018 in three main geographical regions (Latin America and the Caribbean, South Asia and Sub-Saharan Africa according to the definition of the World Bank, which can be compared to European countries as an exemple of mortality trends in HICs - see Figure 1.1).

The increase over the last 60 years has been important, however regional, infra-regional and national trajectories are unequal and improvements still need to be made in comparison to HICs. Latin America is completing its epidemiological transition initiated earlier, with non-communicable disease constituting the main causes of mortality, and a life expectancy at birth surpassing on average 75 years old. On the other hand South Asia and Sub-Saharan Africa are still undergoing major transformations characteristic of the second stage of the health transition model. Our study will focus on the latter two regions.

Whereas South Asia's trajectory appears to be quite linear, the period from the 1980s to the 2000s marked a standstill if not a regression of health improvement in Sub-Saharan Africa, due to the combination of the economic crisis, the HIV epidemics, the relative decline of vaccination campaigns and the resurgence of malaria on account of hydroxychloroquine resistance. This reminds us that the path towards the elongation of life expectancy is not always linear as it may appear with Omran's model, but often includes downturns and periods of stagnation. This continental or subcontinental overview masks national and sub-national disparities. In the case of South Asia's seemingly linear trajectory some countries experienced similar drawbacks as can be observed in Sub-Saharan Africa; Sri Lanka, for example, underwent a slight decrease in life expectancy at birth during the 1980s due to the civil war.

Understanding these evolutions will be key to supporting and keeping up the improvement of health conditions in LMICs. However, as we will see, the lack of data and the complex nature of these transformations represent challenges to researchers and local health systems.

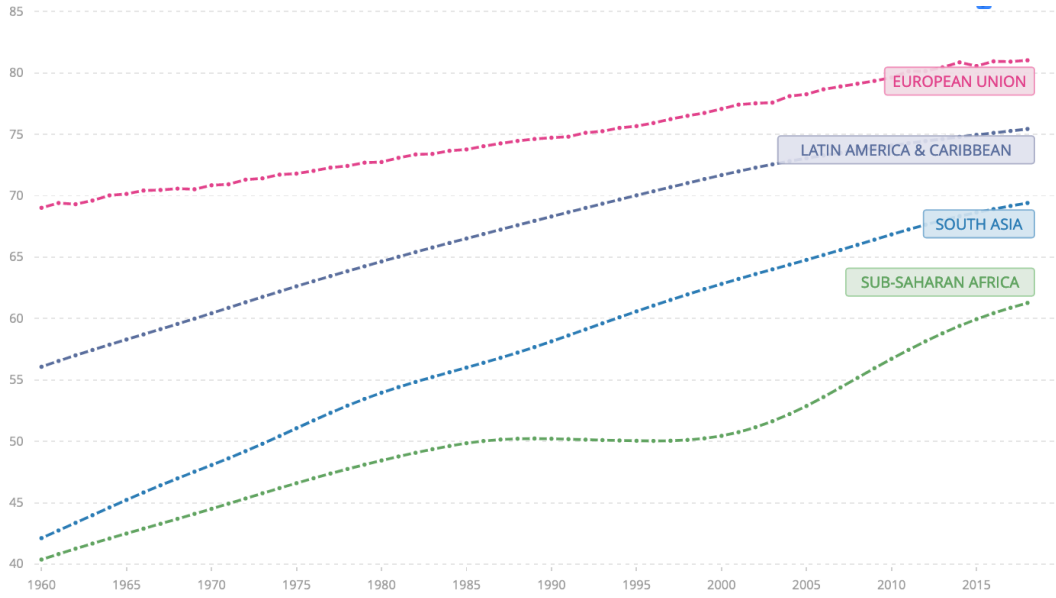


Figure 1.1: Evolution of the life expectancy at birth (in total, years, both sexes) from 1960 to 2018, in Latin America and the Caribbean, South Asia, Sub-Saharan Africa and the European Union

Data from the World Population Prospects, United Nations Department of Economic and Social Affairs, Demographic profiles https://population.un.org/wpp/Graphs/DemographicProfiles/926.
Data vizualisation from the World Bank website : https://data.worldbank.org/indicator/SP.DYN.LE00.IN?end=2018&locations=8S-ZG-ZJ&most_recent_value_desc=true&start=1960&view=chart consulted in January 2021

1.2 Studying the health transition, a statistical challenge

1.2.1 'A scandal of invisibility' : the lack of Civil Registration and Vital Statistics in LMICs

As defined by the United Nations, civil registration and vital statistics (CRVS) systems are a continuous, permanent, compulsory and universal recording of the occurrences and characteristics of vital events (births, deaths and causes of death, foetal deaths, marriages and divorces and other legal requirements defined in each country). They represent a central sources of information for demographers and public health specialists to monitor short- and long-term demographic changes and mortality trends. In particular, the recording of deaths and causes of deaths is the cornerstone of analysing the health transition, enabling to monitor the trends of mortality and causes of death over time.

However, continuously recording complete vital statistics requires an important amount of infrastructure and investment, especially for cause-of-death data as a medical diagnosis is needed. These data are also relatively recent on a large historical perspective ; it is not until the nineteenth century that CRVS start developing at a national scale in Europe and the United States. The recording of causes of death in particular raises the question of the classification of causes of death and its standardisation in first establishing regional and national statistics but also cross-country comparisons (Meslé et Vallin, 1998). This is how the International Classification of Disease (ICD) came into existence during the second half of the nineteenth century to standardise the recording of causes of death [1].

However, half of the world's births and two-thirds (38 million) of 56 million deaths across the world still go unrecorded, as stated by the WHO, with LMICs accounting for the vast majority of these unrecorded vital events. Cause-specific mortality data are even more scarce. Figure 1.2 represents civil registration coverage by country : as

[1] Efforts of harmonisation have been made since the middle of the nineteenth century, with a first attempt to create an international nomenclature of causes of death in 1853 at the International Statistical Congress. But the ICD has been truly operational only since the turn of the century with the third classification of 1893 (Meslé et Vallin, 1998). However, during this period CRVS systems were still in devlopment ; in France, it was not until 1906 that cause of death registration covered the entire national territory.

is illustrated be seen, cause-of-death data are non-existent for Sub-Saharan Africa (except Mauritius, the Seychelles islands and South Africa) and some South Asian countries such as Cambodia, Vietnam, Laos or Burma, and are only partial for countries such as Malaysia, India or even China. The two latter countries have put in place in parallel to the development of their CRVS a system of civil registration through random sampling, allowing them to collect representative data while recording vital events for only a fraction of the population, though with less precision (Rao and Mantal, 2020).

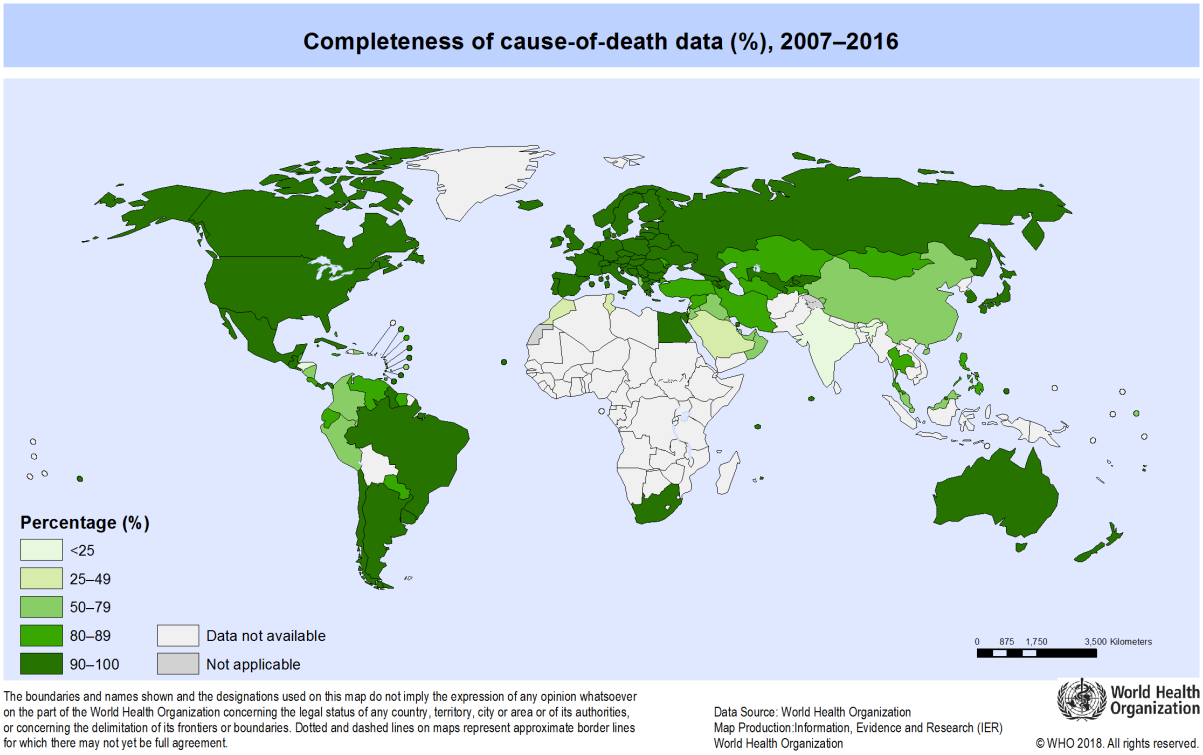


Figure 1.2: Civil Registration coverage of causes of death (%), 2007-2016

World Health Organisation, 2018
Last update of data 2016

This lack of civil registration data seems dire as it not only impedes the monitoring of health and demographic trends, but most of all deprives citizens of an important tool for their rights to be respected. Indeed legal documents are crucial to prove identity and citizenship and thus to provide access to local services and recognition of rights of property and inheritance, and can also represent a protection against some forms of systemic abuse (Setel et al., 2007.).

Even though it has been on the agenda of the United Nations since 1968, CRVS have experienced very little improvement, even through the new impulses carried by the Millennium Development Goals and more recently the Sustainable Development Goals (Mikkelsen et al., 2015). Numbers also reflects who counts ; as stated by Setel et al. in their 2007 paper, this still represents a true "scandal of invisibility" for the world's most fragile population.

1.2.2 Alternative sources of cause of death data : from national estimation models to HDSS (Health and Demographic Surveillance Systems)

However, alternative sources of data exist, even though they should represent only a transitional solution as they are not equivalent to an established and functional CRVS system, which remains the aim of the UN Statistical Office and the WHO.

A wide array of partial alternative sources such as hospital data, additional census questions, national surveys and indirect public policy information can also be used to estimate mortality levels and causes of death. These alternative data are the foundation of estimation models. They provide a helpful tool to estimate number and cause-of-death data, as well as burden of diseases, most frequently at a national level. Today, the Global Burden of Disease project carried out by the Institute for Health Metrics and Estimation (IHME) in collaboration with WHO is the main provider of such estimates for most countries in the world.

These estimations underline the considerable shift in causes of death experienced in the last 40 years in LMICs : the rise of non-communicable diseases as main causes of death, especially with cardiovascular diseases now considered the first cause of death in Sub-Saharan Africa, South Asia and Latin America. However, the quality of the estimations remain dependent on the quality of data available, which as we have seen are very scarce for LMICs. Moreover, the lack of transparency behind IHME models has given rise to criticism from the scientific community (Tichenor and Sridhar 2020 ; Mathers, 2020). Indeed, as models and sources for the estimations are not public, assessment of the assumptions of the models and of the quality of the estimations are impossible. These limits make it very hard to analyse in detail the health transition from this data, especially as the health transition can be taken partly as an assumption in these models [2].

The Health and Demographic Surveillance System (HDSS), on the other hand, represents a unique alternative source of data suitable for more detailed analysis of mortality at a local level. HDSS, called 'population observatories' in French are delimited geographical areas monitored in order to create civil registration and vital statistics by regularly collecting information about vital events (births, deaths, marriages...), also setting a framework for various subject-specific surveys. With the oldest sites dating back to the 1960s, the number of population surveillance systems has been developing, especially since the 1980s, to compensate for data scarcity (Delaunay, 2018). They have been put forth as a particularly robust and interesting source of data to monitor health trends and global goals set by the international community, such as the recent Sustainable and Development Goals for 2030 (Sankoh, 2017).

Recording causes of death has long been one of the aims of HDSS. However, the standard death recording system in HICs - certificates of death elaborated by physicians after examination of the body -, is very costly and not feasible in LMICs settings, especially as an important number of deaths occurs far from health facilities. To overcome these difficulties, a pragmatic, interim solution has been developed in HDSS : verbal autopsies. They are questionnaires, administered to the primary caretaker of the deceased, asking a series of standardised questions about the symptoms and medical history of the deceased. This questionnaire is carried out soon after the death, though after a customary period of mourning varying according to local traditions ; the information collected is then interpreted by physicians, or more recently by algorithms to determine the probable cause of death.

First developed in the seventeenth century in London to monitor epidemics but superseded by systems of death registration in HICs, systematic interviews started to be carried out in the 1950s and 1960s in Asia (Khanna and Narangwal in India, Companiganj in Bangladesh) and in Africa (Keneba in the Gambia) (Garenne and Faveau, 2006). The methodology then spread in HDSSs sites, with Matlab (India) and Niakhar (Senegal) producing their first verbal autopsy questionnaires in the late 1970s early 1980s, to then spread through networks of sites such as the INEDPTH network (for more details see the presentation of the data, section 2.1).

However, beyond the lack of data, the complexification of morbid processes and the possible cumulative burden of diseases brought about by the health transition seem to call for complementary means of investigation to monitor these evolutions.

[2]The independence and objectivity of these estimations have also be quesitoned on the premices of conflict of interest between the IHME, the WHO and the Bill and Melinda Gates Foundation (principal funder of the IHME), as they are both the commissionners of the estimations and evaluate their policies on the basis of those evaluation Tichenor and Sridhar 2020 ; Mathers, 2020). But this criticism is in majority due to the lack of data, as this conflict of interest could be assessed if the models and its assumptions were transparent.

1.3 Healthcare systems and the health transition : the importance of studying multimorbidity to understand the challenges to come

1.3.1 The complexification of the morbid process

The health transition represents an important challenge for health care systems and public policies, first and foremost because of the increasing burden of non-communicable disease. According to the WHO, each year, 15 million people die from a NCD between the ages of 30 and 69 years; over 85% of these "premature" deaths occur in LMICs [3]. This important shift in the demand of health care services represents a double challenge of training and infrastructure (Martini and Figg, 2010). These evolutions have raised the awareness of international organisations, as the fight against NCDs has been listed amongst the Sustainable Development Goals in 2015.

Indeed, health care systems and public policies have been traditionally geared towards the treatment of infectious, neonatal and maternal diseases. And contrary to a large-scale vaccination campaign, most NCDs require routine monitoring, chronic and often very costly treatments. Moreover, NCDs tend to be chronic diseases very susceptible to comorbidities, especially with age (Desesquelles, 2015). In addition, long-term prevention seems key in the battle against NCDs upstream ; lifestyle choices (such as diet, exercise) are indeed determinant factors to their development. Indeed, policies of prevention and education have played a key role in HICs in the progress against cardiovascular diseases in the 1970s.

The dichotomy between infectious diseases and NCDs has to be nuanced, as HIV for example requires both chronic treatment and important investments in prevention. However, the question of the extent of these evolutions remain to be asked, as very few data exist. Most of all, the analysis of the patterns of multimorbidity (risk factors and multimorbidities) appears as a priority to inform public policies

1.3.2 A cumulative burden of disease ?

This necessity to better understand the complexification of morbid processes has long been argued in HICs (Academy of Medical Sciences, 2018), especially as their population is rapidly ageing. The situation of LMICs brings a new perspective to this argument. Indeed, with the rise of NCDs, some countries, especially lower income countries, have experienced a persistence of infectious, neonatal and maternal diseases. This so-called 'cumulative burden of disease' (Boutayeb, 2006) calls for a deeper understanding of the interaction between diseases in this particular context.

Indeed, the question of the precise definition of 'cumulative burden of disease' remains central, as there exist several different interpretations:

- It can be first interpreted as **a coexistence between several categories of disease burden at the population level**, which is undoubtedly true. This seems to have been the meaning intended by Abdesslam Boutayeb in his 2006 paper « The double burden of communicable and non-communicable diseases in developing countries », where he coined the expression. Other burdens have since then been put forth, such as the rise of road injuries, giving way to the more inclusive formulation of 'cumulative' burden of disease. The question then remains whether this situation appears as a logical intermediary step in the process of the health transition or if it could represent a risk of deterioration of the current progress. In light of this, the continuous and detailed monitoring of mortality trends remains very important.

- Second, **the question of a possible cumulative burden of disease at the individual level** remains. As the Covid-19 pandemic has shown, there are multiple possible interactions between communicable and non-communicable diseases: in an important number of cases, individuals suffering from NCDs are more susceptible to infectious disease, and infectious diseases can lead to NCDs. The coexistence of malnutrition in the form of under-nutrition and over-nutrition in Bangladesh for example (Kolčić, 2012) leads to the question of whether a single individual could combine the two through an overly caloric but insufficiently nutritive diet.

[3] Publication from 2018, https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

This cumulative burden would most likely affect the most vulnerable portions of the population in urban areas, where lifestyle changes increase the rate of NCDs. Hence, investigating the following question seems particularly valuable : to what extent can this cumulative burden of disease be observed at the individual level ?

Coined in 1996 by van den Akker and her colleagues (Almirall and Fortin, 2013), the concept of multimorbidity seems to be able to contribute to the answer to these theoretical questions. It has been developped in HICs in a context of an ageing population in order to investigate the co-occurrences of diseases in an individual: having the potential to shed light both on complex mortality processes and on the possible interactions between communicable and non-communicable diseases. It will be the focus of the current report.

1.4 Developing methods to understand and face the challenges of the transition : analysing multimorbidity using an algorithmic interpretation of verbal autopsies (InterVA)

Defined as the study of the occurrence of multiple diseases or conditions occurring in a single individual, multimorbidity has been attracting increasing attention since the beginning of the 2010s (Ford and Ford, 2018), as more emphasis has been put on the fact that patients often have more than one health problems that should not be considered in isolation.

The unique context of health transition in low- and middle-income countries facing both an increasing prevalence of non-communicable disease, and the persistence of communicable disease, seems to give a new perspective to the concept and calls for adequate research. However, even though the idea is being increasingly put forth (Banerjee et al, 2020), studies on the subject pf LMICs remain very limited, especially because of the scarcity of data.

The aim of this project is to contribute to building potential methods and tools to analyse multimorbidity in this context of limited data. After a presentation of the umbrella term that is multimorbidity and its different approaches, we will present our own approach using verbal autopsy mortality data for the Indepth HDSS Network, aiming to take advantage of all the information gathered through verbal autopsies and their algorithmic interpretation through InterVA.

1.4.1 Defining multimorbidity : an approach developed in HICs given a new perspective in the context of LMICs

Despite growing interest in the subject, there is to this day no consensus regarding the precise definition of multimorbidity and how its measure should be carried out, making cross-study comparison difficult. These divergences also explain the enormous discrepancies in the evaluation of the prevalence of multimorbidity : a review of prevalence study in HICs found estimates ranging from less than 15% to more than 70% in the general population largely due to the differences is the operational definition used (Fortin and al. 2012.). The only constant results seems to be that multimorbidity increases with age (with a general increase from 40 onwards to hit a plateau in the 70s, although these conclusions are mainly drawn from data from HICs- Furhman, 2014)

First and foremost, multimorbidity is in most cases **clearly differentiated from comorbidity**. Comorbidity (term coined in 1970 by Alvan R. Feinstein) studies prevalence and weight of different co-occuring clinical conditions for a designated index disease that could be either a cause or a consequence of this index disease. For example, hypertension, obesity, non-alcoholic fatty liver disease or sleep apnea are common comorbidities for patients with diabetes. On the other hand, **multimorbidity aims to be more general : the term is used when no index disease is under study**. Hence a patient suffering from diabetes and non-alcoholic liver disease has a multimorbidity (Fuhrman, 2014 ; Johnston et al, 2019).

However, studies have been using diverse definitions of multimorbidity to study these co-occurrences. The following question summarises the main divergence between definitions : **Which conditions should be taken into account ?** This is described in the following points :

- **How detailed should the list of included disease be ?**
  The main divergence between definitions come from which conditions are taken into account, hence which cases are considered as multiple diseases and which are not. The debate can be summarised by how detailed the list of diseases should be. Indeed, the aim of most studies is to focus on the relations between what could be called **"cluster of diseases" or "patterns of multimorbidity"** (Prados-Torres et al. 2014) : to detect common associations of diseases that appear non-random, and in doing so isolate common disease clusters that could inform physicians on common co-morbidities to look for and stimulate research to define treatment targeted towards these associations of diseases rather than the juxtaposition of single-disease treatments.
  These cluster analysis can be done at varying degrees of details, from identifying large general cluters (eg. in Prados-Torres et al. 2014, cardiovascular disease and metabolic disorders, mental health clusters and muscoskeletal disorders) to common associations of diseases with their ICD-10 categories. In this context, the question of how complications of diseases should be considered, an important point that is however rarely addressed in the definitions (Johnston et al, 2019).
  On the other end of the spectrum, some definition aim to **"include conditions more than just diseases**[4] (Almirall and Fortin, 2013), in order to take into account a broader spectrum of indicators of patient health, in particular mental health, but also in some rarer cases socioeconomic factors.

- **Should multimorbidity focus on risk factors more than just diseases ?**
  This second point is closely related. With this aim to broaden the definition, some studies treat risk factors as diseases, such as hypertension, alcohol consumption, diet, weight, etc... It can also include some symptoms such as chronic fatigue, migraines, back pain, visual impairment or urinary incontinence (Willadsen and al., 2016) to name a few. This broadened definition can take two forms : the study of the combination of risk factors and the study of the combination of risk factors and diseases. This seems justified insofar as these risk factors are the main basis for action for patients and physician. However the relevance of comparing the analysis of risk factor co-occurrences (especially without linking them to patient pathologies) with a multimorbidity exclusively focused on diseases can be put into question, but can vary depending on the risk factors considered as some risk factor are also diseases (such as diabetes, obesity depending on the definition of disease/condition).

- **Should chronicity of the condition(s) be a criteria ?**
  Another main divergence between definitions is the consideration it gives to chronic diseases. Indeed, some studies only consider co-occurring chronic conditions in the study of multimorbidity. However, a majority of studies referring to this criteria, require that only one of the two diseases to be chronic, considering the co-occurrence of diseases with a certain list of chronic diseases.
  This seems sensible in a context where complications of diseases do not want to be taken into account, and in a perspective of understanding the long-term risks of chronic diseases. But it can be more questionable if the aim is to isolate possible clusters of diseases. This also lead to the following question, what should be the **specification of the duration of the condition** to be considered chronic ? (ex : more that X amount of time, or occurred in the last Y amount of time)

All in all these differences between definitions **depend on the purpose of the study** (and of course the available data) : from a purely clinical perspective -aiming to isolate disease clusters-, to a perspective in well-being trying to estimate the impact on quality of life, with public health perspectives lying in between, and trying to gather information on how to prevent this health burden.

However, researchers and clinicians seem to agree that it is **a concept geared towards epidemiologists and researchers** not towards clinicians and patients (Willadsen et al, 2016). It indeed seems to have little concrete operability for physicians on individual basis, but rather emphasises for them the necessity of a patient-based rather than condition-based treatment.

[4]The difference between "medical condition" and "disease" seems hard to establish. We could attempt a definition with the following sentences : *a **disease** would be a defined entity based on either symptoms or objective measures, or both* ; *a **condition** on the contrary would refer to a broader concept defining both diseases, isolated symptoms or a socioeconomic situation*

1.4.2 Measuring multimorbidity : an overview of existing methods

The different measures of multimorbidity found in literature exemplifies the divergence in the definition. These measures can be divided in two main categories : index base measures and disease counts.

Indexes of multimorbidity : estimating individual and societal effect

Some studies use indexes to measure morbidity, such as the Charlson Index, the Cumulative Illness rating, the ACG System to name a few (Johnston et al 2019). In general, these indexes have the particularity of focusing more on **measuring the burden of multimorbidity and its consequences, rather than the nature of multimorbidity in itself**. For example, the Charlson Index estimates the 10-year survival of patients from their existing conditions[5]. Some focus more on the consequences of multiple diseases or symptoms for an individual patient, such as the Duke Severity of Illness Checklist, providing a scale of the burden of illnesses from all identified health problems[6]. Other focus more on the consequences on societies and the health system, such as the Adjusted Clinical Group (ACG) System[7], aiming to evaluate the health costs for certain subsets of the population considering their specific clinical characteristics. The source of data are varied, ranging from clinical data, to all survey representative of nation-wide population.

All in all, these measures seem to result from a very broad definition of multimorbidity, aiming to be more directly operational for health care professional, but less relevant in an epidemiological perspective.

Disease count : a clinical and epidemiological perspective

The second type of measure is more closely related to the epidemiological perspective : counting the number of co-occurring diseases (or/and risk factors) according to the definition of conditions taken into account (as discussed earlier). There are some discrepancies on the cut-off point of the number of diseases that should be considered multimorbidity, but a great majority of the literature seems to agree on the simple criteria of two or more conditions.

Diverse sources of data

The sources of data are also very diverse : nationally representative surveys, local multimorbidity-focused surveys, and clinical data, as a non-exhaustive list. There are three main sources of measure of conditions : self-reported, biomedical measures, or physician diagnosis.

Multimorbidity measure in LMICs, a quick review

Developed in HICs, the interest of multimorbidity in LMICs has also been rising to analyse the increasing burden of non-communicable diseases and population ageing; however, studies remain scarce and the literature calls for further developments (Pati et al., 2015; Banerjee et al. 2020), especially in Sub-saharan Africa, as most current study to this day have been based in South Asian countries.
Most studies until now rely on self-reported conditions through surveys, allowing for more important sample population, with a smaller amount of literature based on physician diagnosis (Pati et al., 2015). They underline the importance of chronic conditions and the need for their integration in health policies and care, especially hypertension, diabetes, asthma and skin conditions (Pati et al., 2015), as well as psychological and emotional problems (Zhang et al., 2019). However, as in HICs, definitions of multimorbidity vary considerably between studies making comparison difficult.

[5] Available on the following website : https://www.mdcalc.com/charlson-comorbidity-index-cci
[6] For the details of the index see https://doi.org/10.1007/978-94-007-0753-5
[7] Details available at http://mchp-appserv.cpe.umanitoba.ca/viewConcept.php?printer=Y&conceptID=1304

1.5 Our approach : estimating multimorbidity at stake in mortality through the identification of multiple causes of death

1.5.1 The project : exploring the potentiality of local cause-specific mortality data in LMICs from verbal autopsies

This summary underlines the necessity of clearly defining and justifying the definition of multimorbidity chosen in each study. Considering the scarcity of data in the context of LMICs and the novelty of the approach of multimorbidity in this context, we decided to start from the available data.
Most studies exploring multimorbidity in LMICs to this day use clinical trial data. Based on medical and biological analysis, they offer very precise and good quality data, however those clinical surveys are very costly and remain scarce, and moreover they allow only to survey small samples of a specific population. On the other hand, as we have seen, some studies take advantage of large population health surveys to estimate multimorbidity based on individual declaration. Providing information on a larger scale but with a poorer quality of data, they seem complementary to clinical studies.

However, it appeared that **very few studies have been done focusing on the burden of multimorbidity in mortality**. Even though multimorbidity represents an important burden in terms of quality of life that should not be reduced to its impact on mortality, evaluating the burden of multimorbidity in mortality appears as an important starting point to direct further research and raise awareness on the question. Moreover, even though mortality data remain scarce in LMICs, due to the frailty of the CRVS, representative verbal autopsy data are routinely collected at a local level in HDSS sites and remain available. Elaborating tools to study multimorbidity on these data would allow the development of a potentially generalisable approach to routinely monitor multimorbidity, and would be of great value to keep track of heath trends in the context of the sanitary transition.

To do so, we were inspired by the multiple causes of death approach developped by the Multicause Network[8] for the analysis of morbid processes in HICs. Indeed, the study of mortality traditionally takes into account only underlying causes coded by the physician according to the rules set by the WHO in the International Classification of Diseases. However, the recent development of the multiple causes of death approach has shown the value of analysing all coded causes of death (associated and underlying) to understand the morbid processes leading to death Désesquelles 2010, 2012 et 2016, Barbieri et al. 2017). In particular, this approach has underlined the importance of considering associated causes, as their omission leads to important underestimation of the burden of certain diseases, especially chronic conditions such as diabetes. More restrictive than a general approach of multimorbidity, it considers only conditions possibly coded as causes of death as potential co-morbidities; a clear definition easy to replicate across studies.

In this study, we aim to investigate if we could adapt this approach to the cause-specific mortality data available in LMICs, namely verbal autopsies, in particular when a probabilistic algorithm of interpretation is used to determined the causes of death.

1.5.2 Taking advantage of the algorithmic determination of causes of death through InterVA : a potentially generalisable approach to estimate multimorbidity

To undertake this endeavour we decided to settle for a particular form of verbal autopsies (VA) and tool to determine causes of death. Indeed, developed since the 1980s, several standards of VA exist in HDSS sites, mostly depending on the seniority and the particular history of the sites. However, since 2007, the WHO has been developing VA standards, aiming for a harmonisation of questionnaires and of interpretation of causes of death, with regular updates of these standards (in 2013, 2014 and 2016 respectively). This uniformisation went hand in hand with the increasing use of algorithmic tools to determine causes of death that have been in development since the early 2000s, providing a very cost-effective and standardised substitute for physician interpretation. The standardisation of these algorithmic interpretation of causes of death seems particularly interesting, as it offers a framework to generalise potential successful methods.

[8] https://mcod.web.ined.fr/wiki/Accueil

Several such algorithms are used today, such as the Tariff method or InSilicoVA[9], and others are still in development. However, to this day, one of the most widely used algorithms of VA interpretation is InterVA. Developed by Peter Byass and his team at Umea university in Sweden (http://www.byass.uk/interva/crms) since the early 2000s (Byass, 2003), it is based on a Bayesian approach and has the particularity of determining up to three causes of death per case. Associating each cause with a probability, it was mainly conceived to estimate the burden of different causes of death at a population level.

Our aim will be to evaluate whether it could also be used to estimate multiple causes of death at an individual level, in order to analyse the role of multimorbidity in mortality.

1.5.3 An exploration to be taken as a proof of concept

However, we should underline the limits of this projects. First of all, the selection bias of verbal autopsy data should be kept in mind. Indeed, compared to clinical data or death certificates available in HICs, verbal autopsies remain poor-quality data, as only information collectable by laymen from the caretakers of the deceased, sometimes a few months after the death occurred, is available.

But most importantly, even though the idea seem to bear great potential, it remains an exploratory project, where we explicitly use data outside of its original intended scope. Indeed, the question of the accuracy of the interpretation of cause of death is already being debated among the scientific community. Hence, attempting to measure different forms of multimorbidity from this data could seem to some very far-fetched, as it was not designed towards this purpose. **This project is to be taken more as a proof of concept : an exploration of the feasibility of approaching multimorbidity through verbal autopsy data**.

2. Data and sources

2.1 Cause-specific mortality data interpreted by InterVA-4 from 22 HDSS sites from the INDEPTH Network

The International Network for the Demographic Evaluation of Populations and Their Health (INDEPTH) is a network of 49 HDSS sites in 20 countries across Africa, Asia and the Pacific region, monitoring over 3.5 million people. Created in 1998, it aims to provide support and coordinate HDSS sites in regards to their methods, data collection and data analysis, in order to provide high-quality longitudinal data in LMICs[1]. One of the objectives of INDEPTH has been the standardisation of methods, in order to insure greater comparability across sites. InterVA, developed inside INDEPTH, contributed to this harmonisation concerning mortality data collection and analysis.

More than 72,300 deaths of adults across Sub-Saharan Africa and Asia

Our database benefits from this harmonisation : made available by INDEPTH, it consists of **cause-specific mortality data from 22 HDSS sites members of the INDEPTH Network (INDEPTH, 2014) elaborated using a standardised methodology : verbal autopsies interpreted by InterVA-4**. Elaborated to enable comparison across sites (Streadfield et al, 2014 (1)), it provides deaths by sex, age groups, sex, cause(s) of death and their associated probability. Note that the results are not necessarily nationally representative. **We selected deaths of adults (15 years old or more) to focus on the question of multimorbidity**.

This database gathers 72 330 deaths captured through VA (89.6% of the deaths that occurred), from 22 sites across 13 countries in Africa (14 sites and 76% of deaths) and 4 countries in Asia (8 sites and 24% of deaths) (Table 2.1). With only two urban sites accounting for 4% of deaths, the vast majority of deaths occurred in rural or rural dominant settings. Years of death range from 1992 to 2012, though 89% of deaths occurred after 2003, a date that marks an increase in the number of HDSS sites. As InterVA-4 was released in 2012 on the basis of the 2012

[9]For an overview and analysis of current algorithms see McCormick et al, 2015.
[1]For more information see http://www.indepth-network.org

WHO VA standard, this means that the available VAs were all retrospectively transformed into the WHO 2012 and InterVA-4 input format for processing. Among the deaths of adults, 39% occurred before 50 years old, 20% between 50 and 64 and 41% after 65. Regarding sex, 49% of the deceased are female, 51% are male.

Table 2.1: Deaths according to geographic and demographic characteristics, INDEPTH

	Frequency	Percentages
Deceased by sites		
Continent		
Sub-saharan Africa	55079	76.1
Asia	17251	23.9
Country		
South Africa (2 sites)	18331	25.3
Kenya (3 sites with 1 urban)	17633	24.4
Bangladesh (4 sites)	13568	18.8
Ghana (2 sites)	10891	15.1
Burkina Faso (2 sites with 1 urban)	3515	4.9
India (2 sites)	2214	3.1
The Gambia (1 site)	1603	2.2
Malawi (1 site)	1367	1.9
Senegal (1 site)	1036	1.4
Indonesia (1 site)	775	1.1
Vietnam (1 site)	694	1.0
Cote d'Ivoire (1 site)	375	0.5
Ethiopia (1 site)	328	0.5
Urban/rural		
rural	69504	96.1
urban	2826	3.9
Deceased by demographic characteristics		
Age group		
15-49 years	28005	38.7
50-64 years	14414	19.9
65 + years	29911	41.4
Sex		
female	35350	48.9
male	36980	51.1
Year of death		
1992-2002	8131	11.2
2003-2007	33864	46.8
2008-2012	30335	41.9

From 72 330 VA of adults in 22 sites of the INDEPTH HDSS Network, 1992-2012 data

The important burden of infectious and parasitic diseases

Table 2.2 illustrates the distribution of causes of death, determined using the method prescribe by InterVA of weighing each cause by its associated probability (see infra for the detailed method). Among adult death, 42% of deaths are attributed to "diseases of poverty", almost entirely accounted for by infectious and parasitic diseases. The share of non-communicable diseases is almost equivalent, with 39% of deaths, whereas accidents and injuries account for a much smaller percentage of deaths (6%).

Table 2.2: Cause-specific mortality fraction according to the probabilities interpreted by InterVA

Cause	Frequency	Percentages
« Diseases of poverty »	**30341.43**	**41.95**
Infectious and parasitic diseases	29176.42	40.34
Anemia and malnutrition	515.72	0.71
Maternal CoD	648.29	0.90
Neonatal CoD	1.00	0.00
Non-communicable diseases	**28197.27**	**38.98**
Cancers	9020.21	12.47
Diabetes and cardiovascular diseases	11795.35	16.31
Chronic respiratory diseases	2578.12	3.56
Other non-communicable diseases	4803.59	6.64
Injuries and violent deaths	**4558.13**	**6.30**
Indeterminate	**9233.17**	**12.77**

From 72 330 VA of adults in 22 sites of the INDEPTH HDSS Network, 1992-2012 data

This important share of infectious and parasitic diseases is quite surprising considering that the population of interest consists of adult deaths only, where a higher share of non-communicable diseases could have been expected. This exemplifies the concept of the cumulative burden of disease that remains very important for adults, especially in rural Africa[2].

2.2 The detailed VAs the HDSS of Ouagadougou, Burkina Faso, 2010-2019

More than 1,700 VAs of all ages

To supplement this cause-specific mortality database, we used detailed data from 1,714 VAs from the Ouagadougou HDSS, Burkina Faso, to test our approach with the detailed information present in VAs but that are not available for the INDEPTH database. We considered VAs from all age groups in order to analyse as many VAs as possible as our number is already limited. These VAs were collected from 2010 to 2019, with 2010 to 2012 being a test period for InterVA's verbal autopsy questionnaire in the Ouaga HDSS (representing 157 VAs, 9% of our dataset), and 2013 marking the start of all VAs done with this questionnaire, with 870 deaths (50%) occurring from 2013 to 2015 and 687 deaths (40%) from 2016 to 2019 (Table 2.3).

With children under 15 representing almost a third of the deceased (30.4%), more than four in ten (43.6%) were older than 50 at the time of death. There are also more men than women among the deceased, with 55% men versus 45% men.

An urban HDSS with a heavier burden of non-communicable diseases

An interesting feature of the HDSS of Ouagadougou is its urban context, being set in neighborhoods on the outskirts of the capital of Burkina Faso. This particularity is translated in the epidemiological profile of its population with a lower burden of diseases of poverty and higher burden of non-communicable diseases, as can be seen in Table 2.4 presenting the CSMFs of the site. Non-communicable diseases are attributed the highest mortality fraction in Ouagadougou, with more than half of adult deaths (51%) attributed to non-communicable diseases compared to 39% in average for the 22 HDSS sites included in the INDEPTH database. On the other hand, the burden of "diseases of poverty" estimated for adults is responsible for less than a third of deaths (31%), compared to 42% in average in the INDEPTH database where they remain the leading causes of death. This context appears especially interesting when studying multiple causes, as non-communicable diseases tend to lead to comorbidities.

[2]For a detailed analysis of the distribution of causes of death across sites see Streatfield et al. 2014 (1).

Table 2.3: Deaths according to demographic characteristics, HDSS of Ouagadougou

	Frequency	Percentages
Age group		
< 5 years	390	22.8
05-14 years	131	7.6
15-49 years	446	26.0
50-64 years	265	15.5
65 + years	482	28.1
Sex		
female	776	45.3
male	938	54.7
Year of death		
2010-2012	157	9.2
2013-2015	870	50.8
2016-2019	687	40.1

From 1,714 VA of all age groups from the HDSS of Ouagadougou, 2010-2019

Considering all age groups however, the burden of diseases of poverty is higher, with an attributed 43% of deaths, comparable to the percentage of death attributed to non-communicable diseases (38%), and injuries and accidents account for a much smaller amount of death (6%).

Table 2.4: Cause-specific mortality fraction according to the probabilities interpreted by InterVA, Ouagadougou

Cause	Freq.	%	Freq. adults	% adults
« Diseases of poverty »	**729.32**	**42.55**	**364.95**	**30.59**
Infectious and parasitic diseases	587.77	34.29	332.01	27.83
Anemia and malnutrition	26.38	1.54	12.54	1.05
Maternal CoD	20.76	1.21	20.40	1.71
Neonatal CoD	94.41	5.51	NA	NA
Non-communicable diseases	**658.56**	**38.42**	**607.88**	**50.95**
Cancers	134.70	7.86	133.53	11.19
Diabetes and cardiovascular diseases	302.91	17.67	287.30	24.08
Chronic respiratory diseases	37.36	2.18	35.23	2.95
Other non-communicable diseases	183.59	10.71	151.82	12.73
Injuries and accidents	**100.18**	**5.84**	**82.64**	**6.93**
Indeterminate	**225.94**	**13.18**	**137.53**	**11.53**

From 1,714 VA of all age groups (adults >= 15) from the HDSS of Ouagadougou interpreted by InterVA-4, 2010-2019

2.3 A detailed analysis of InterVA-4's algorithm

As we aim to assess InterVA's potentialities and limitations, InterVA-4's model itself also constitutes one of our main sources.

We took advantage of the open source nature of the software to study the way it processes data. We analysed in particular its core formula and its a priori matrix that constitutes the basis of the model. Called probbase, it is available as part of the downloadable software from the InterVA website[3] or from the openVA package available

[3] http://www.byass.uk/interva/crms

on Github or CRAN [4]. Key articles helped us in this endeavour, notably the presentation of the InterVA-4 model (Byass, 2012) and the critical analysis of the model by McCormick and his co-authors (McCormick, 2015).

We also benefited from fruitful exchanges with colleagues of the late creator of InterVA-4 Peter Byass, who contributed to its elaboration, namely Edward Fottrell and Lucia d'Ambruoso.

3. Sources and specific research question : identifying multiple causes of death with InterVA-4

The idea of this project is to take advantage of a particular feature of InterVA, the fact that it selects up to three causes per death, to investigate if it allows us to identify multiple causes of death.

However, we have to take into account that the causes determined by InterVA are of a very different nature than the data collected from death certificates filled by physicians that are used in HICs to analyse multiple causes. This investigation requires therefore extra care and a detailed analysis of the nature of the data used and of the processing of InterVA to be able to access to what extent multiple causes can be identified. This is the object of the present section.

3.1 InterVA, a tool to interpret verbal autopsies

3.1.1 The objective : monitoring causes-specific mortality in contexts of limited data and resources

Since the 2000s, several model of automatic interpretation of verbal autopsies have been developed and evolved into several methods still used today (see Leitao et al, 2013 for more detail). InterVA is one of the most widely used of those models and has been in development since 2003 (Byass, 2003) by Peter Byass and his team at Umea University in Sweden and is regularly updated (the last version InterVA-5 was released in 2020 with a module targeted to diagnos Covid-19 deaths). It was designed to determine cause specific mortality statistics in the most efficient and cost-effective manner, to inform local authorities in a context of limited mortality data and resources.

Developed inside of the INDEPTH HDSS Network, the idea behind this tool is to provide an automatic and replicable method to interpret available data : verbal autopsies, information collected via questionnaires from the parents or primary caretakers of the deceased about the events and the symptoms leading to death. Based on an algorithm and on the standardised VA questionnaire regularly updated by the WHO (aligned to the WHO 2012 Verbal Autopsy instrument in the case of InterVA 4 (Byass, 2012)), the objective of InterVA is to offer an alternative to physician interpretation of VA that has the advantages of being a lot more cost-effective, easily replicable and providing comparable interpretations through the different geographical and epidemiological contexts across low and middle income countries.

The aim of the algorithm is to promote the implementation of VA and their interpretation and facilitate the monitoring of deaths and causes of death at a population level, in settings where it would otherwise not be possible. To do so, it provides for each death up to three probable causes each with an associated probability, that are to be are summed up to estimate the cause-specific mortality fraction of the population over the given period. It is this process of determination of causes that we are going to detail in the following section.

3.1.2 A Bayesian model...

Determining causes of death from verbal autopsies comes down to the following question : given the symptoms and characteristics reported in the VA, can we deduce the cause of death ? InterVA's model takes a probabilistic approach to this question : **given the symptoms and characteristics reported in the VA, can we deter-**

[4] https://openva.net/

mine the probability of each specific cause being the cause of death ? And in doing so, what cause or causes appear the most probable ?

To do so, the algorithm is based on Bayes' theorem of conditional probability, determining the probability of an event A occurring given that the event B is true ($\mathbb{P}(A|B)$), chich can be written as follows :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A) \cdot \mathbb{P}(B|A)}{\mathbb{P}(B)}$$

Where :

- $\mathbb{P}(A|B)$ is the probability of A occurring given that B is true,
- $\mathbb{P}(A)$ the unconditional probability of A, $\mathbb{P}(B)$ the unconditional probability of B,
- $\mathbb{P}(B|A)$ the probability of B occurring given that A is true.

InterVA takes VA and causes of death as an application of this formula. Considering a predetermined classifications of all possible causes of death $C_1, ..., C_N$, we can deduce the probability of the particular cause of death C_i given the set of symptoms and characteristics of the deceased declared in the VA $S = \{s_1, ..., s_M\}$ called indicators ($\mathbb{P}(C_i|S)$) as follows :

$$\mathbb{P}(C_i|S) = \frac{\mathbb{P}(C_i) \cdot \mathbb{P}(S|C_i)}{\mathbb{P}(S)} = \frac{\mathbb{P}(C_i) \cdot \mathbb{P}(S|C_i)}{\sum_{k=1}^{N} \mathbb{P}(C_k) \cdot \mathbb{P}(S|C_k)}$$

Hence we can determine the probability of C_i causing the death, if we know for all possible causes C_k the unconditional probability of this cause $\mathbb{P}(C_k)$, i.e. its prevalence in the population, and the probability of presenting the set of symptoms reported S, given that the cause of death is C_k ($\mathbb{P}(S|C_k)$).

In order to frame the problem in the simplest way, all indicators are considered as binary variable : $s_j = 1$ if the symptom or characteristic j is reported in the VA, otherwise $s_j = 0$. This means that characteristics that comprise several category correspond to several binary indicators ; there is for example 7 indicators corresponding to the 7 age groups categorised by InterVA. Moreover, all numerical (mainly indicators of duration of symptoms) are dichotomised. This can be illustrated with the example of the information collected regarding fever, which corresponds to three independent indicators : "fever of any kind", "fever lasting more than two weeks or more", "fever lasting less than two weeks".

At this point, in order to make the problem more tractable, **InterVA makes a first simplifying assumption that, given a cause, all indicators are independent**. Hence, $\mathbb{P}(S|C_i) \approx \prod_{j=1}^{M} \mathbb{P}(s_j|C_i)$. This is in reality impossible, as particular symptoms often go hand in hand, and it is illustrated by the filters on the questionnaire : questions, hence possible reported indicators, are asked according to age group and sex among other characteristics. But it seems to be a necessary simplification as it would be very impractical and incredibly time-consuming to determine the probability of each combination of 245 symptoms. This assumption is however considered reasonable for most cases by other experts (McComick et al, 2015), even though it most likely discards valuable information.

Hence, the probability of C_i given the set of indicators reported in the VA $S = \{s_1, ..., s_M\}$ can be rewritten as follows :

$$\mathbb{P}(C_i|S) = \frac{\mathbb{P}(C_i) \cdot \mathbb{P}(S|C_i)}{\sum_{k=1}^{N} \mathbb{P}(C_k) \cdot \mathbb{P}(S|C_k)} \approx \frac{\mathbb{P}(C_i) \cdot \prod_{j=1}^{M} \mathbb{P}(s_j|C_i)}{\sum_{k=1}^{N} \mathbb{P}(C_k) \cdot \prod_{j=1}^{M} \mathbb{P}(s_j|C_k)}$$

InterVA makes a second simplifying assumption : only present indicators, i.e. $s_j = 1$, are taken into account in the computation. One the one hand, this assumption seems reasonable in a context of important uncertainty. Indeed, for an adequate diagnosis, distinguishing between the absence of a symptom and the lack of knowledge regarding this symptom might be crucial. However, given that the information is reported by a third party that may not be able to make the difference, it might be wiser to only take into account symptoms that are

obviously present, and discard all other information. If necessary, it is possible to create an additional indicator pertaining to the absence of a symptom that appears especially important. This is the case for example for an eventual malaria test that corresponding to two indicators : "recent positive test for malaria", "recent negative test for malaria", a value of 0 for both implying the absence of testing for malaria or ignorance thereof.

On the other hand, this assumption appears as the principal limitations of InterVA, that the model *InSilicoVA* for example aims to overcome (McCormick, 2015). This assumption not only implies disregarding precious information about the absence of certain symptoms, in practice most of the data collected, but also means that the computed probabilities are not comparable across individuals, making it is impossible to construct a valid measure of uncertainty (McCormick, 2015). This feature of the model remains up for debate.

Hence, the final formula used by InterVA can be written as follows :

Let J be the set of present indicators : $j \in J \Leftrightarrow s_j = 1$

$$\mathbb{P}(C_i|S) = \frac{\mathbb{P}(C_i) \cdot \prod_{j \in J} \mathbb{P}(s_j|C_i)}{\sum_{k=1}^{N} \mathbb{P}(C_k) \cdot \prod_{j \in J} \mathbb{P}(s_j|C_k)}$$

Hence, the InterVA makes the assumption that we can deduce the probability of cause given the symptoms declared ($\mathbb{P}(C_i|S)$) if we know the probability of all possible causes $\mathbb{P}(C_k)$, and the probability of presenting each individual symptom s_j given that the cause of death is C_k ($\mathbb{P}(S|C_k)$).

From there on, we only need to estimate those probabilities to determine the probability of the cause of death, arguably the most difficult part of the modelling process.

3.1.3 ... based on an a priori probability matrix elaborated by a panel of experts...

Estimating those probabilities is not a straight forward task. Many automated methods if interpretation (for example: Flaxman et al.,2011; James et al., 2011; Murray et al., 2011a) rely on a so called "gold standard", a database of large number of deaths with causes of death certified by physicians and considered reliable. However, creating those databases is difficult and expensive, and they are mostly collected in hospital settings that can bring important biases where most deaths occur at home (King and Lu, 2008, McCormick, 2015). Moreover, as the prevalence of diseases changes over place and time, either "gold standards" databases can only be used locally, or they would have to contain enough deaths through time and space to be valid across geographical contexts. The goal of InterVA was to develop a method that did not rely on a gold standard.

In that context, the team InterVA made the choice to **consult a panel of experts in order to evaluate these probabilities, asking physicians for to estimate the tendency of observing each indicator given a particular cause of death**.

The estimation process proceeds as follows. The group of physicians make their estimations according to a letter-grading system that is given qualitative interpretations, as can be seen in Table 3.1. These letters are then converted into numbers using a logarithmic scale. Indeed, some experiences have shown that the perception of qualitative expressions of probability tends to correspond more to a logarithmic than a linear scale (Ohnishi, 2002).

These estimations result in a matrix of associating to each of the N causes the probability of presenting each of $M + 1$ indicators (i.e. the M characteristics and symptoms reported in the VA plus the unconditional probability of the cause).

However, as we have mentioned, setting an a priori prevalence (or unconditional probability) can appear as an important limit. As we have already established, these prevalences are likely to vary greatly in time and space, depending on the cause. For this reason, **special arrangements exisst for three causes of death : HIV/AIDS, malaria and sickle cell disease**; their *a priori* probability is set by the user according to the context of the data. The user has to chose between three possible settings : High (= B in Table 3.1), Low (= C) and Very Low

Interpretation	Letter	Value
Always	I	1.0
Almost always	A+	0.8
Common	A	0.5
	A-	0.2
Often	B+	0.1
	B	0.05
	B-	0.02
Unusual	C+	0.01
	C	0.005
	C-	0.002
Rare	D+	0.001
	D	0.0005
	D-	0.0001
Hardly ever	E	0.00001
Never	N	0

Table 3.1: Conditional probability scale : from qualitative frequency interpretations to values (Byass, 2012)

(=E), for HIV/AIDS and Malaria before computing the interpretation. The probability of sickle cell disease is then considered the same as for malaria.

Other causes of death are also likely to vary greatly among geographical settings. However, it is considered that their signs and symptoms are sufficiently specific as to not require an adjustment of their *a priori* probability to the local context (Fottrell et al., 2011). Indeed, as it has been shown by Fottrell et al. (Fottrell et al. 2011), applying changes up to three steps on the logarithmic scale to a random selection of up to 50% of the *a priori* probabilities, leads to the same public health conclusion, with very similar cause specific mortality fractions, and similar ranking of causes. Hence, the model is quite robust to small changes in the a priori probability matrix : it is not necessary to seek absolute precision on each probability, but rather overall plausibility. In this context, adjusting the prevalence of each cause of death was not deemed necessary, as it is only one indicator among 246 for InterVA-4.

3.1.4 ... selecting up to three causes per death and their associated probability

Using the probability matrix and the Bayesian formula written above, InterVA determines the probability of each cause of death. To make this information easier to analyse, InterVA then selects up to three causes per death according to the following rules (even though the probabilities associated to each cause is also available as an output of the model) :

- If no cause has an associated probability of at least 0.4, then the cause of death is considered indeterminate with probability 1,
- Otherwise, the cause with the highest probability p_1 is assigned as the first cause of death alongside its probability,
- A second cause is selected if the second highest probability p_2 represent at least half of the highest probability:: $p_2 \geq 1/2p_1$
- And a third cause is selected if the third highest probability p_3 represents at least half of the highest probability: $p_3 >= 1/2p_2$

Any residual probability (i.e. $1 - (p_1 + p_2 + p_3)$) is considered as a partial indeterminate.

These causes and their associated probabilities are then used to compute cause-specific mortality fractions at the population level. For each cause (including indeterminate), their associated probabilities are summed, and then divided by the size of population to determine cause-specific mortality fractions that include an indeterminate fraction. Hence one death can account for several causes, and frequencies are usually not round numbers.

3.2 Research question: can we identify multiple causes of death through InterVA ?

Once the inner workings of InterVA have been established, to what extent can the deaths assigned more than one cause of death be interpreted as deaths from multiple causes ?

3.2.1 Close to 11% of deaths assigned more than one cause of death

As Table 3.2 shows, in our data from the INDEPTH Network, 10.7% of deaths are assigned more than one cause of death through InterVA. These percentages appears relatively stable across sites, and according to sex, with only minimal variations by age groups (with a slight increase from 9% for 15 to 49 to 12% for 65 and over).

However, deaths that are associated with three causes only represent 0.5% of the total number of deaths. This is probably due to the fact that the rule to select a third cause is very strict. **As this proportion is very small, we decided to only consider the first two causes of death in our analysis**.

Table 3.2: Deceased by number of causes, INDEPTH

	Frequency	Percentages
One cause	**64596**	**89.3**
More than one cause	**7734**	**10.7**
2 causes	7389	10.2
3 causes	345	0.5

From 72 330 VA of adults in 22 sites of the INDEPTH HDSS Network, 1992-2012 data

At first glance, this percentage, though relatively modest, could account for multiple causes. It does not seem to conflict with current estimation of multimorbidity in LMICs (Pati et al. 2015 meta-analysis of multimorbidity shows estimates that range from 4% to 21% of the population using varying methods and definitions). This seems especially true as multiple causes corresponds more to stricter definitions of multimorbidity, as a lot of risks factors that can be taken into account in multimorbidity are not considered causes of death.

However, at this stage, we cannot insure that those causes represent multimorbidity. Moreover, the important stability of the percentage of multiple causes identified might lead us to believe it is a structural consequence of the rules of selection of causes. Indeed, we do not see a significant increase in the prevalence of multiple causes identified by InterVA in sites where they could be expected. For example, the urban sites of Ouagadougou (Burkina Faso) and Nairobi (Kenya), where the prevalence of NCDs are higher, do not present significantly higher percentages of death associated with more than one cause.

3.2.2 Differentiating between competing and co-occuring causes of death

Table 3.3 presents among deaths with more than one cause of death the most frequent associations of causes in order of frequency. **We decided here to focus our attention on the associations of causes irrespective of their order in regards to their associated probabilities, as we are first and foremost interested in possible associations of causes**. Moreover, order of causes does not seem particularly discriminative : when taken into account, the frequencies of one order versus the other are close for most of associations.

This gives us a first appreciation of the nature of those multiple causes interpreted by InterVA. **Two different mechanism of association seem at play here :**

- **Associations that appear as plausible multiple causes of death, or "co-occurring causes"**. Indeed, in theory, InterVA is capable of detecting a combination of causes leading to death, if indicators of those causes are sufficiently present in the reported symptoms. Moreover, associations such as diabetes mellitus and stroke

appear as an entirely plausible combination of causes leading to death, as they are very documented comorbidities. Likewise, pulmonary tuberculosis is known to be amongst the most common opportunistic infections affecting HIV patients especially in Sub-Saharan African, hence it seems that the association HIV/AIDS and pulmonary tuberculosis could be interpreted as multiple causes of death.

- **Associations that seem to result of an inability to decide between two possible causes of death that are mutually exclusive, or "competing causes"**. On the other hand, InterVA was primarily designed to estimate cause-specific mortality at the population level in a context of limited information. The bayesian approach and the possibility to select up to three causes was also designed to handle this uncertainty by allowing to estimate an indeterminate fraction, but also selecting several causes of death weighed by probabilities when it is difficult to decide between two different diagnostics due to a lack of information. The association of acute respiratory infection including pneumonia and malaria seem to pertain to this logic. Indeed, pneumonia and malaria are very difficult to distinguish without the help of a biomedical test (Källander, 2004), hence it appears difficult to interpret this association without complementary information ; it appears reasonable to suppose that this association results more from an inability to decide between those two causes than a multimorbidity. The same interpretation seems to hold for the association of tuberculosis and respiratory neoplasm, as well as reproductive and digestive neoplasms, as an important number of VA indicators rest on the localisation of pain, but also for associations such as road accident and assault, as the intention behind an incident is not easily capture in closed VA questions.

Note that, at this stage, we do not distinguish between underlying, immediate and intermediate causes of death, as defined by the International Classification of Disease. Indeed, as much as those distinctions are useful, our data is very far from presenting a sufficient amount of information to make these distinctions. By gathering all these different natures of causes under one label, "co-occuring causes", we aim first and foremost to separate those multiple causes that can be interpreted as multimorbidity, with different underlying processes of combination, from associations of causes that were not both present at the moment of death, only resulting from a lack of information.

Hence, a question remains : **taking into account the way InterVA operates, could we differentiate co-occuring causes from competing causes ?** The following sections will take you through the method we propose to do so and its results, to then discuss its possible uses and limitations.

4. Our method: using a similarity index to distinguish co-occurring from competing causes of death

4.1 An overview of the method : hypotheses, concepts and limitations

4.1.1 Identifying the associations of causes with a very similar symptomatology

The aim of our project is to determine if we can differentiate "co-occurring causes of death" - associations of causes that we can interpret as multiple causes of death - from "competing causes" - associations resulting from an inability to decide between two plausible causes of death that are mutually exclusive, i.e. that were not both present at the moment of death.

To do so we formulated the following hypotheses :

- **Associated causes with very similar symptoms and demographic characteristics, i.e. that have very similar associated probabilities in the *a priori* probability matrix, are most likely competing causes**. For example, malaria and acute respiratory infections have very similar symptoms, and are hard to differentiate without the help of a malaria test, hence we would tend to interpret them as competing causes.

Table 3.3: Most frequently associated causes of death by order of frequency

Cause A	Cause B	Frequency	Percentages
Pulmonary tuberculosis	HIV/AIDS related death	392	5.07
Acute resp infect incl pneumonia	Sepsis (non-obstetric)	339	4.38
Acute cardiac disease	Stroke	319	4.12
Pulmonary tuberculosis	Respiratory neoplasms	309	4.00
Reproductive neoplasms MF	Digestive neoplasms	268	3.47
Acute resp infect incl pneumonia	Malaria	208	2.69
Acute cardiac disease	Other and unspecified cardiac dis	184	2.38
Acute resp infect incl pneumonia	Pulmonary tuberculosis	179	2.31
Digestive neoplasms	Other and unspecified neoplasms	179	2.31
Diabetes mellitus	Stroke	148	1.91
Road traffic accident	Assault	148	1.91
Acute resp infect incl pneumonia	Other and unspecified cardiac dis	141	1.82
Pulmonary tuberculosis	Other and unspecified cardiac dis	123	1.59
HIV/AIDS related death	Intentional self-harm	117	1.51
Acute abdomen	Digestive neoplasms	115	1.49
Other and unspecified cardiac dis	Chronic obstructive pulmonary dis	115	1.49
Pulmonary tuberculosis	Chronic obstructive pulmonary dis	109	1.41
Acute abdomen	Diarrhoeal diseases	103	1.33
Stroke	Other and unspecified cardiac dis	100	1.29

Associations irrespective of order, with frequency >= 100

From 7,734 VA of adults with more than one cause of death in 22 sites of the INDEPTH HDSS Network, 1992-2012 data

- On the other hand, **causes associated to very different symptoms are most probably co-occurring causes as their association is highly unlikely to result from a confusion of the two causes**. They are on the contrary most probably capturing the presence of a set of two different groups of symptoms and can therefore reasonably be interpreted as co-occurring causes.

Hence, by quantifying how similar two causes are according to InterVA's definition of the causes (i.e. the *a priori* probability matrix), we could distinguish causes with similar symptoms from causes with sufficiently different symptoms. This would allow us to discard from our analysis associations of causes with similar symptoms that we deem too likely to be competing causes, and analyse as multiple causes only associations of causes that we deem to have symptoms sufficiently different so that we are confident in interpreting them as co-occurring causes.

Note that this is a restrictive approach : we will not be able to identify co-occuring causes with very similar symptoms. If the two causes have very similar symptomatology, i.e. the symptoms reported in the VA are associated to very close probabilities and do not appear specific enough, it is impossible to deduce if only one or both diseases lead to death. Indeed, a double infection of pneumonia and malaria could lead to death, but without biomedical tests information, it is impossible to know if a person was suffering from one or the other, or even both conditions. In this context, we exclude associations that could result from confusion, keeping in mind this limitation, rather than analysing associations that could result from an undecisive diagnosis

4.1.2 An identification that can be carried out at different levels depending on the amount of detail present in the dataset

Once these hypotheses are established, two main questions remain : **How to quantify the similarity between causes, i.e. how to calculate the similarity index ? And from this index, how to choose the threshold differentiating competing and co-occurring causes ?**

These are the questions we aim to answer in the following sections. First, we will present our core approach, then test its robustness with two variations adapted to two different levels of detail that may be present in mortality data interpreted by InterVA-4 :

- **First, we will present a general approach, based on InterVA-4 *a priori* probability matrix as a whole**, to identify the possible competing associations of causes based on how each cause is defined in the probability matrix. This will allow us to estimate probable competing causes from any cause-specific mortality data interpreted by InterVA-4 without having access to the detailed information reported in the VA, as in the INDEPTH mortality dataset.
- We will then aim to choose a threshold to differentiate between competing and co-occurring causes.

- **Finally, we will present a more specific approach, based on the individual indicators reported in the VA of each death**. This will allow us to identify on a case-by-case basis the risks of confusion between causes, especially for contexts with very limited information where risks of confusion between causes are high. To do so, we will use the detailed VA information of the HDSS of Ouagadougou (Burkina Faso), from 2010 to 2019. This will further allow us to test the reliability of our first general approach, and to assess to what extent the detailed information of VAs are necessary to assess multiple causes of death.

4.2 A general approach : calculating indexes of similarity from the *a priori* probability matrix

4.2.1 The Euclidean and absolute norm indexes of similarity

For InterVA, all possible causes of death (listed in the Appendix, Table 9.1) are defined by a probability vector : a vector associating a given probability to all the indicators possibly reported in the VA, in the case of InterVA-4, to all possible 246 indicators. **Therefore, assessing the similarity between two causes comes down to assessing how similar their probability vectors are : how close are the probabilities associated to each cause, for each indicator ?**

Formally, we can do so by calculating the distance between the two vectors, through a norm function. Hence, we decided to create a similarity index ranging from 0 to 1, by calculating the distance between the two vectors normalised by their sum. Formally, it can be written as follows, given A the probability vector associated to cause a, and B the probability vector associated to cause b, the similarity index between a and b is :

$$I_{a,b} = I_{b,a} = \frac{\parallel A - B \parallel}{\parallel A + B \parallel}$$

where for a given vector X, $\parallel X \parallel$ is a norm function. Therefore, the more similar the causes, the closer the index is to 1, and the more different the causes, the closer the index is to 0. We applied this formula to all possible associations of causes using two different norm functions :

- The Euclidean norm, defined as follows : $\parallel A - B \parallel = \sqrt{\sum_{i=1}^{246} [\mathbb{P}(s_i|a) - \mathbb{P}(s_i|b)]^2}$,
- The absolute norm : $\parallel A - B \parallel = \sum_{i=1}^{246} |\mathbb{P}(s_i|a) - \mathbb{P}(s_i|b)|$

with $\{s_1, ..., s_{246}\}$ the 246 indicators taken into account by InterVA-4. These two functions are very similar ways to estimate distance between two vectors, with only slight differences, that allow us to test the robustness of our results.

To compute all the indicators, we used the default probability matrix of InterVA-4 that has both malaria and HIV set to "very low". The index could be computed for all possible settings, and hence tailored to the settings of each site. However, as we will argue in the discussion, this does not lead to any significant changes in the indexes, as a priori prevalence is only one value among 246, and thus we decided to use one single probability matrix for more clarity.

4.2.2 Two highly correlated indexes with similar distribution

As Figure 4.1 and Table 4.1 illustrate, the two indexes have very similar distributions. Considering all possible associations of causes defined by InterVA-4, the Euclidean index ranges from 0.28 to 0.91, with a median at 0.71. The absolute norm index has slightly lower values and a slightly more spread out distribution, with values ranging from 0.14 to 0.85 and a median at 0.63.

Figure 4.1: Distribution of the similarity indexes (InterVA4)

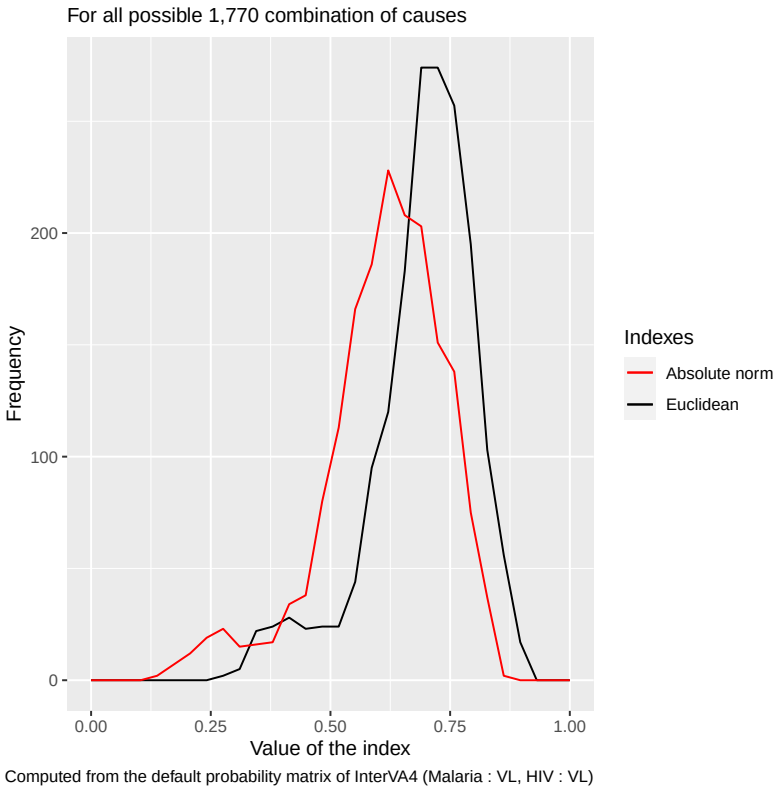
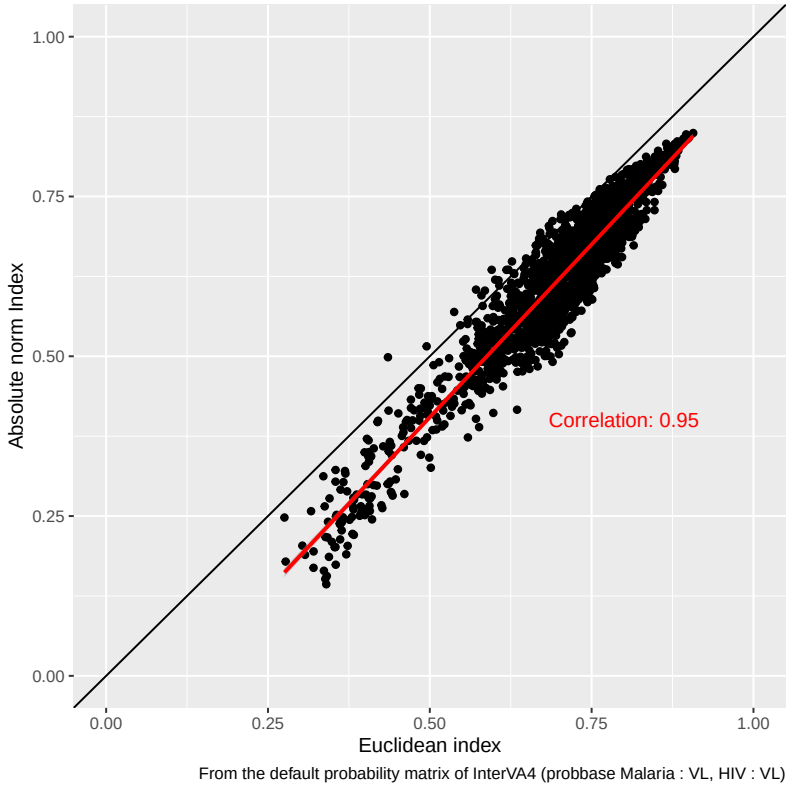


Table 4.1: Distribution of the indexes of similarity : Median, Mean and Quartiles

	Euclidean index	Absolute norm index
Min.	0.28	0.14
1st Qu.	0.65	0.55
Median	0.71	0.63
Mean	0.69	0.61
3rd Qu.	0.77	0.70
Max.	0.91	0.85

As we could expect, the two indexes are highly correlated, with a Pearson correlation coefficient of 0.95 (Figure 4.2), with the value of the absolute norm index being in majority inferior to the value of the Euclidean index for a given association of causes. Hence, they both convey a very close appreciation of how similar two causes are in regards to their symptomatology and demographics. However, a slight difference should be underlined : **given their definition, the absolute norm index gives more weight to small differences in probability compared to the Euclidean index**, where the emphasis is set on important differences in probability. This difference explains the occasional discrepancy between the two indexes on some associations of causes.

Figure 4.2: Correlation between the indexes of similarity



4.2.3 Mapping the possible confusions between causes

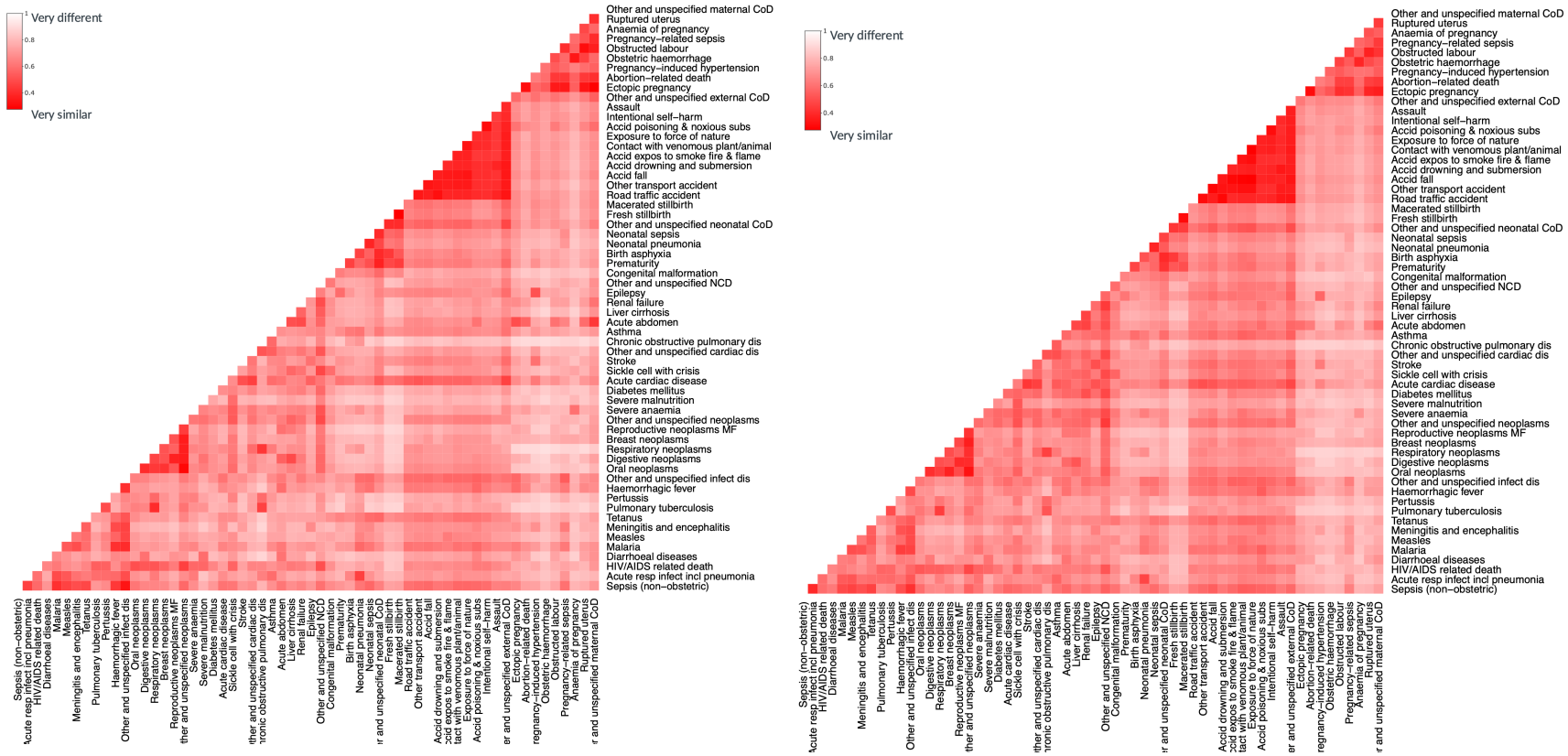
From these indexes, we can draw a heat map of possible confusions between causes according to the *a priori* probability matri (Figure 4.3). With on the right-hand side and on the bottom all causes of death, we visualise all 1,770 possible associations of causes and their index. The redder the index, the more similar the two associated causes, i.e. the risk that their association results from competition rather than a co-occurrence is higher[1].

We can see that certain causes of causes stand out as being particularly different from other causes, especially when considering a group of causes as a whole : for example, maternal as well as neonatal causes of death appear particularly distinct from the other groups of causes, which was to be expected. Some individual causes also stand out as particularly distinct, i.e. characterised by a light colour : this is notably the case for chronic obstructive pulmonary diseases, pulmonary tuberculosis, malnutrition and diarrhoeal diseases.

On the other hand, within groups of causes, we can see high levels of similarity : this seems especially true among cancers - where the localisation of the cancer is not always easy to establish from a VA-, maternal causes of deaths or external causes of death (accidents and violent deaths) - where the intentions behind the event are not always easy to disentangle with closed questions. Moreover, some causes stand out as particularly similar, making their associations probable competing causes : malaria and acute respiratory disease including pneumonia, as we could have expected, but also malaria and sepsis, respiratory neoplasm and pulmonary tuberculosis among others.

[1] As the index is symmetrical, we presented only the bottom half of the map. And as we are focused on the possible associations of causes, we did not represent the diagonal (the associations of each cause with itself that have an unsurprising similarity index of 0).

Figure 4.3: Maps of possible confusions between causes



(a) Heatmap of the Euclidean index of similarity

(b) Heatmap of the absolute norm index of similarity

Among the possible confusions, those with HIV/AIDS are interesting to note. Indeed, according to InterVA-4, **HIV/AIDS related deaths and pulmonary tuberculosis are very similar causes**, ranking amongst the 20% most similar associations for both indexes (see Table 4.3 infra). Indeed, after analysing the most probable symptoms of HIV/AIDs related deaths as defined by the *a priori* matrix, possible opportunistic infections are taken into account in the definition of HIV (see Annex, Figure9.2). This is understandable, as the indicators of HIV alone are very scarce. However, because of this definition, we unfortunately cannot interpret tuberculosis and HIV as multimorbidities using this method, as the risk of confusion is quite high. This difficulty echoes to the wider question of the difficulties of identifying HIV in the VA already pointed out in the analysis at the population level of HIV mortality (Streatfield, P. Kim et al. 2014. (2) and (3)).

Hence, more than similar symptomatologies identifiable by physicians, **the indexes allow us to point out structural similarities in the definition set by InterVA especially when they take into account possible opportunistic infections or comorbidities**. This is especially important as the status of comorbidities in InterVA definitions is not clearly defined, as the question was not central to the creation of the tool intended to determine population level of cause-specific mortality.

4.3 Determining a threshold to distinguish co-occurring from competing causes

4.3.1 Confronting the indexes with the INDEPTH database : the importance of competing causes

These indexes allow us to have a general overview of the similarities between causes based on the definitions set by InterVA. **However, from these 1,770 possible associations of causes, only 323 are found in the data**, with varying degrees of frequency (see Annex Figure 9.1 for a heatmap illustration in the reduction of causes to consider). Indeed, many combinations are highly unlikely and some even impossible because of demographic constraints : still birth could not be combined with maternal causes for example. Confronting these indexes with the associations found in the data will allow us to investigate with more precision the possible and most frequent risks of confusion between causes.

Table 4.3 presents the most frequent combinations of causes confronted with the indexes (the exhaustive table is available upon request). As the raw value of the index is hard to interpret on its own, it is more helpful to appreciate its place in the distribution of all possible causes. Here we used the deciles : the smaller the decile the more similar the causes are according to its index of similarity. And as we can see, **among the most frequent associations of causes, most appear very similar, their index of similarity belonging to the first or second decile**. This could have been expected, as most of these associations of similar causes appeared already as probably resulting from confusion such as pneumonia and malaria, acute cardiac diseases and stroke or associations between neoplasms. However, some such as HIV/AIDs related deaths and tuberculosis underline one main limitation of our method : the inability to identify multimorbidity with similar causes due to the limited information in our data.

This important prevalence of probable competing causes is also illustrated by the distribution of the indexes amongst the associations present in the data, as showed by Table 4.2 : the median of the Euclidean index of the associations present in the data is 0.65 compared to 0.71 in all possible associations, and for the absolute norm index it is 0.52 compared to 0.61. Indeed, the frequency of higher index values is lower in the data, which seems understandable as it is higly unlikely for an important number of these very different causes to be associated in real life.

Table 4.3 also shows that, even though the two indexes are highly correlated, they differ in some individual cases. For example here, the association of diabetes mellitus and stroke is similarly appreciated by the two indexes : according to the Euclidean index, they are quite different causes, ranking amongst the 60% most different associations of causes. On the other hand, according to the absolute norm index, they appear as rather similar causes, ranking amongst the 20% of most similar association of causes. This is due to the fact that the absolute norm puts more weight on smaller differences in probabilities, whereas the Euclidean norm will first and foremost assess the high differences in probabilities. Hence, as diabetes mellitus and stroke have different main symptoms (See Annex Figure 9.3 and 9.4 for the list of the most important indicators for those two causes according to InterVA),

they have some similarity amongst certain indicators, especially the less important ones, as they are both causes affecting the cardiovascular system. This particularity leads to a different appreciation of their similarity by both indexes and plays in both directions, some associations are deemed more similar by the Euclidean index compared to the absolute index (such as malaria and acute abdomen, the exhaustive tables of associations and their indexes can be provided on request). This discrepancy between both indexes, however infrequent, can inform our selection of a threshold to isolate probable co-occurring causes.

Table 4.2: Distribution of the indexes of similarity of associations present in the INDEPTH data

	Euclidean index	Absolute norm index
Min.	0.28	0.19
1st Qu.	0.56	0.46
Median	0.65	0.54
Mean	0.62	0.52
3rd Qu.	0.70	0.61
Max.	0.85	0.79

From the 323 associations of causes detected by InterVA-4 in the INDEPTH data

4.3.2 Selecting a threshold

At this stage, it is important to chose a threshold to distinguish competing and co-occurring causes. Several possibilities seem viable and marginally the threshold will remain arbitrary but is essential for any analysis.

First of all, as both indexes are highly correlated, the choice of one index over the other is not particularly crucial. **We decided to use the Euclidean index, as giving more weight to important discrepancy between probabilities seemed the best strategy to identify causes with different sets of main symptoms**, even though the absolute norm index is slightly more spread out in its distribution. However, most results appear robust to the use of the absolute norm, except, as we have already underlined, cases like the association of diabetes mellitus and Stroke.

Choosing a threshold, however, seems a more arduous task, as no clear discontinuity emerges from the distribution. In this configuration, it seemed more suitable to choose a threshold according to quantiles. **Using quartiles, the threshold of** 0.65 **appeared particularly interesting as it allows us to select as probable co-occurring causes 75% of all possible associations of causes and half of the associations present in the data, while selecting combinations that appear sufficiently dissimilar to be interpreted as probable multimorbidity** (the exhaustive table is available upon request). Figure 4.4 illustrates where this threshold stands according to the different sets of associations of causes that can be considered : first, all 1,770 theoretically possible associations of causes, then the 323 associations of causes present in the data (18% of all possible associations), and finally the distribution according to all deaths with more than one cause (7,734 individuals). As we will discuss further in the results section, **this threshold identifies 20.5% of deaths with more than one cause as "co-occurring" causes**. The associations considered as probably competing, according to this threshold, are illustrated in Figure 4.5. We can see that within groups, associations are mainly removed from the analysis of multiple causes, especially associations of cancers, violent death and maternal deaths that, as we saw, have particularly similar symptoms, as well a number of associations with HIV/AIDs, particularly neoplasms and respiratory infections.

Table 4.3: Associated causes confronted to the similarity index

Cause A	Cause B	Frequency	Euclidean index	Euclidean decile	Absolute norm index	Absolute norm decile
Pulmonary tuberculosis	HIV/AIDS related death	392	0.55	1	0.48	2
Acute resp infect incl pneumonia	Sepsis (non-obstetric)	339	0.46	1	0.28	1
Acute cardiac disease	Stroke	319	0.50	1	0.33	1
Pulmonary tuberculosis	Respiratory neoplasms	309	0.41	1	0.35	1
Reproductive neoplasms MF	Digestive neoplasms	268	0.37	1	0.29	1
Acute resp infect incl pneumonia	Malaria	208	0.49	1	0.45	1
Acute cardiac disease	Other and unspecified cardiac dis	184	0.46	1	0.36	1
Acute resp infect incl pneumonia	Pulmonary tuberculosis	179	0.59	2	0.51	2
Digestive neoplasms	Other and unspecified neoplasms	179	0.36	1	0.29	1
Diabetes mellitus	Stroke	148	0.71	5	0.52	2
Road traffic accident	Assault	148	0.38	1	0.22	1
Acute resp infect incl pneumonia	Other and unspecified cardiac dis	141	0.73	6	0.61	5
Pulmonary tuberculosis	Other and unspecified cardiac dis	123	0.73	6	0.63	6
HIV/AIDS related death	Intentional self-harm	117	0.63	3	0.53	2
Acute abdomen	Digestive neoplasms	115	0.56	1	0.50	2
Other and unspecified cardiac dis	Chronic obstructive pulmonary dis	115	0.52	1	0.42	1
Pulmonary tuberculosis	Chronic obstructive pulmonary dis	109	0.47	1	0.43	1
Acute abdomen	Diarrhoeal diseases	103	0.61	2	0.53	3
Stroke	Other and unspecified cardiac dis	100	0.61	2	0.48	2

Associations of causes irrespective of order, with frequency > 100

From 7 734 AV of adults with more than one cause of death.

4.4 Testing the robustness of this approach with detailed VA data (HDSS of Ouagadougou, Burkina Faso)

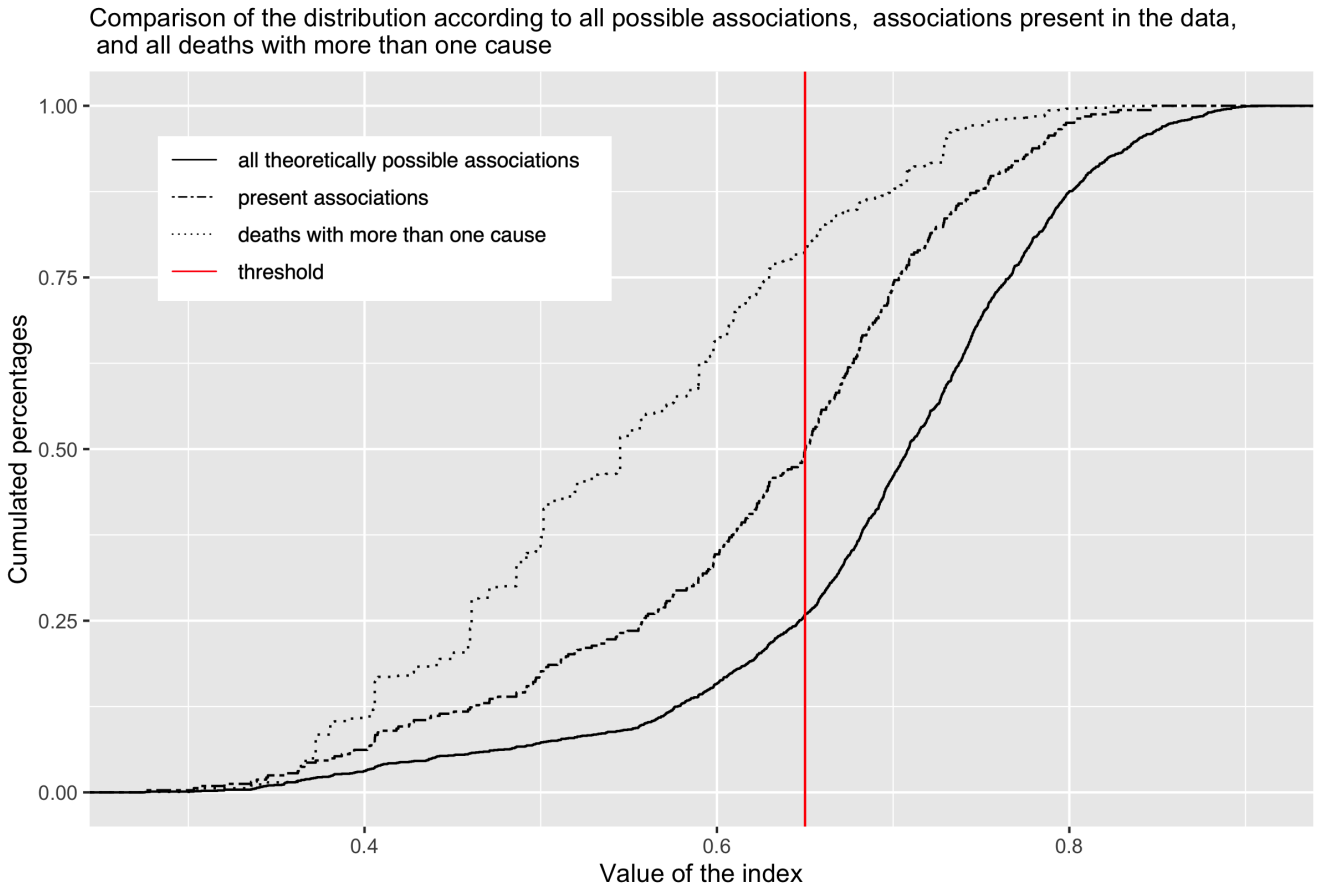
4.4.1 A broad but limited approach : the limited information in VAs confronted to all possible indicators

The approach presented above provides a critical analysis of the 246 (indicators) by 60 (causes) matrix of *a priori* probability of InterVA to identify the causes with very similar definitions.

However, as we have pointed out, to determine the cause of death, InterVA only takes into account indicators reported in the VA. And as Table 4.4 shows, the average number of indicators reported in a VA is far below the 245 possible indicators[2]. **In average, only 21 indicators are reported**, in part due to the fact that considering gender and age filters (around 60 indicators can be reported given the age group and the sex of the individual), but also because of the lack of information. As we can see, 25% of the VAs from the Ouagadougou HDSS report only

[2]245 indicators + 1 a priori prevalence of the cause of death, hence 246 probability can be taken into account by InterVA-4 for all death. In Table 4.4, we have not taken into account the a priori prevalence as an indicator: to give an example for the individual that has reported 3 indicators, 3 + the a priori prevalence, hence 4 *a priori* probabilities are taken into account to compute the probable cause(s) of death.

Figure 4.4: Comparaison of the cumulative distributions of the Euclidean similarity index



Interpretation : 25% of all possible causes have an associated Euclidean similarity index inferior to 0,65, it is the case for 50% of the associations present in the INDEPTH database. Finally, considering all deaths attributed more than one cause, 80% are associated to an index inferior to 0.65

14 indicators, with an important number of questions unanswered, leaving the algorithm with little information to make a diagnosis.

Taking into account only 21 indicators to compute the probabilities of each cause of death may change the similarity between causes in both directions. If none of the distinctive indicators is reported in the VA, two a priori rather different causes could be very similar or completely identical. On the other hand, if only distinctive indicators are reported, two causes a priori quite similar could in fact be quite distinct according to this individual's symptom.

However, the detailed VA data are more difficult to access, and are not available for example for the INDEPTH database. **The aim of the present section is hence to present a second approach, which we will call the "case-by-case approach" or the "empirical index" to take into account the fact that InterVA only considers reported indicators. We will then compare it to the general approach in order to assess its robustness considering this limitation.** To differentiate the two approaches we will call the index presented in this first approach the "theoretical index".

Figure 4.5: Competing and co-occurring causes according to the selected threshold

All associations with an Euclidean index of similarity inferior to 0.65 are considered competing

4.4.2 Computing an individual index according to the symptoms reported in each VA

To test this hypothesis, we elaborate an individual index applying the same formula but taking into account only the indicators effectively reported in each VA[3]. Formally, it can can be written as follows.

Given $S = (s_1, ..., s_M)$ the vector of indicators reported in the VA - with $s_j = 1$ if the indicators is reported in the VA, and $s_j = 0$ if it is not reported in the VA[4] -, A the probability vector associated to cause a, and B the probability vector associated to cause b. We use the element-wise vector product, also known as the Hadamard product, noted $\odot$ where $(A \odot B)_{ij} = (A)_{ij} \odot (B)_{ij}$. The empirical similarity index between a and b is :

$$I_{a,b} = I_{b,a} = \frac{\parallel A \odot S - B \odot S \parallel}{\parallel A \odot S + B \odot S \parallel}$$

[3]Hence, taking into account the indicators declared in the VA out of 245 possible indicators plus the a priori prevalence.
[4]As it is always taken into account, we consider that the indicator for the a priori prevalence is always set to 1

Table 4.4: Number of indicators reported per VA

	Number of indicators
Min.	3.0
1st Qu.	14.0
Median	19.0
Mean	21.3
3rd Qu.	27.0
Max.	61.0

From 1714 VA, HDSS Ouagadougou, Burkina Faso, 2010-2019

Given $\mathbb{P}(s_j|a)$ the *a priori* probability of presenting the indicator j given cause a,

$$(A \odot S)_j = \mathbb{P}(s_j|a) \odot (s)_j = \begin{cases} \mathbb{P}(s_j|a), & \text{if } s_j = 1 \\ 0, & \text{if} s_j = 0 \end{cases}$$

Hence, let J be the set of reported indicators, with $j \in J \Leftrightarrow s_j = 1$, the Euclidean empirical similarity index is :

$$I_{a,b} = I_{b,a} = \frac{\sqrt{\sum_{j \in J}[\mathbb{P}(s_j|a) - \mathbb{P}(s_j|b)]^2}}{\sqrt{\sum_{j \in J}[\mathbb{P}(s_j|a) + \mathbb{P}(s_j|b)]^2}}$$

Each individual with more than one cause is attributed an index ; this means that one particular association of causes can be associated to different index values according to the specific symptoms declared by each individual. This is illustrated by Table 4.5, which presents by decreasing frequency the maximum, minimum, average and standard deviation of the individual index, which is calculated on the 186 deceased attributed more than one cause of death among the 1,714 VAs from HDSS of Ouagadougou [5] confronted with the theoretical index of the same association. This appears to be an asset to this approach, making it possible to distinguish more precisely cases where limited information leads to uncertainty from other cases where two different sets of symptoms corresponding to two different and probably co-occurring causes are identifiable.

We can first see that for some associations of causes, the empirical indexes are all very close, such as sepsis and acute respiratory infections including pneumonia (with an empirical index of 0 for all 12 occurrences, meaning that given the reported symptoms, the two causes are exactly identical), or digestive and reproductive neoplasms (with an index around 0.21, very similar causes). On the other hand, some causes have indexes with an important range, for example acute cardiac disease and stroke (ranging from 0.46, quite similar, to 0.11, very similar), or stroke and acute abdomen (ranging from 0.69, seemingly very different causes, to 0.40 rather similar).

Moreover, there is a substantial difference between the empirical and theoretical indexes. As shown in Figure 4.6, **even though their correlation remains quite high (Pearson coefficient 0.71), the value of the empirical index of similarity is on average lower, and very rarely surpasses that of the theoretical index** : a majority of the points are situated below the y = x axis, and very few points are situated above. **This shows that the limitations of the information reported in VAs where there is more possible confusion, compared to the general definition of the causes relying on the 245 indicators**, but does not necessarily discredit our general approach. First of all, as the indexes are not computed from data with the same order of magnitude, the comparison is not easy, as mechanically differences appear greater when few probabilities are taken into account.

The empirical index seems, in the vast majority of cases, to lead to the same conclusion in regards to the associations considered as competing through the general approach (theoretical index <0.65), as few associations seem to empirically surpass the theoretical index so as to reach a very high empirical index value while having a very low theoretical value. Most importantly, the theoretical index seems to identify total confusion, as illustrated in the

[5] Considering we already have very limited number of deaths with more than one cause, we did not seperate adults (15 and older) from children.

Table 4.5: Empirical vs. theoretical Euclidean index

Association of causes	n	max	min	average	standard deviation	theoretical index
Sepsis (non-obstetric) / Acute resp infect incl pneumonia	12	0.00	0.00	0.00	0.00	0.46
Acute cardiac disease / Stroke	10	0.46	0.11	0.26	0.12	0.50
Acute resp infect incl pneumonia / Malaria	10	0.43	0.24	0.37	0.06	0.49
Acute cardiac disease / Other and unspecified cardiac dis	9	0.46	0.13	0.38	0.10	0.46
Digestive neoplasms / Reproductive neoplasms MF	5	0.25	0.19	0.21	0.02	0.37
HIV/AIDS related death / Pulmonary tuberculosis	5	0.47	0.35	0.40	0.04	0.54
Acute resp infect incl pneumonia / Pulmonary tuberculosis	4	0.59	0.37	0.49	0.09	0.59
Other and unspecified cardiac dis / Asthma	4	0.53	0.43	0.47	0.04	0.55
Stroke / Acute abdomen	4	0.69	0.40	0.57	0.14	0.66
Acute abdomen / Other and unspecified external CoD	3	0.49	0.45	0.47	0.02	0.61
Diabetes mellitus / Stroke	3	0.73	0.56	0.62	0.09	0.71
Diarrhoeal diseases / Malaria	3	0.66	0.58	0.63	0.04	0.66
Malaria / Acute abdomen	3	0.66	0.62	0.63	0.02	0.62
Neonatal pneumonia / Neonatal sepsis	3	0.00	0.00	0.00	0.00	0.38
Stroke / Other and unspecified cardiac dis	3	0.56	0.31	0.40	0.15	0.61

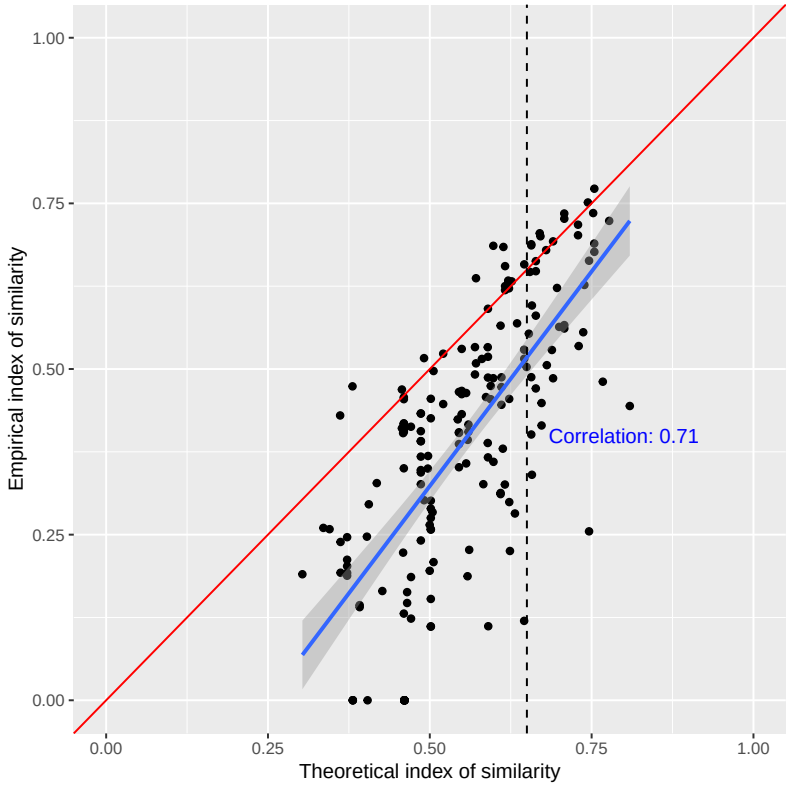
From 186 VA with multiple causes of death determined by InterVA4, Ouagadougou HDSS, 2010-2018

example of sepsis and acute respiratory infectious , with an index of 0, meaning no reported distinctive symptoms between the two causes.

On the other hand, it seems that some associations we might have identified as probably co-occurring seem more competing when detailed symptoms are taken into account. While not a majority, these isolated competing causes identifiable only through their detailed VAs, underline the limitation of our general approach. An analysis of the discrepancies between the two indexes could probably contribute to enriching and modulating the criterion for our general approach ; however, the limited number of the detailed VA data that we have available does not allow us to draw decisive conclusions and calls for the analysis of more detailed VAs.

Hence, this second approach appears to be an interesting alternative to estimating multiple causes of death, when detailed data are available. This calculation would seem especially relevant as part of the algorithm, if a feature to analyse multimorbidity were added. However, an appropriate threshold different from the one of the general approach should probably be selected, considering the difference in the number of probabilities taken into account for computation.

Figure 4.6: Correlation between the indexes empirical and the theoretical Euclidean indexes of similarity



5. Results: the cumulative burden of disease and the burden of NCDs in the multiple causes identified

5.1 2.2% of multiple causes

According to the threshold selected to differentiate co-occurring from competing causes, amongst the INDEPTH mortality database, we identify 1,591 deaths with multiple causes, that is 20.6% of causes attributed more than one cause of death and 2.2% of all deaths (Table 5.1, Table 5.2). It can appear as a relatively small proportion, however, as we will argue in the discussion, our approach of multiple causes is an approach of multimorbidity to be taken as a lower bound considering the limits of the available data.

Table 5.1: Co-occurring causes of death identified through the similarity index, INDEPTH

	Frequency	Percentages
co-occurring causes	1591	2.2
competing causes	6143	8.5
unicause	64596	89.3

From 72 330 VA of adults in 22 sites of the INDEPTH HDSS Network, 1992-2012 data

As can be seen on Table 5.2, **the percentage of co-occurring causes identified increases with age**, from 1.3% of all deaths for 15 to 49 years old to 2.95% for 65 and above. This is not only due to the fact that the percentage of deaths attributed more than one cause is higher with age, the percentage of co-occurring causes among deaths with more than one cause also increases from 15% to 24% between those two age groups. Indeed, we

know that multimorbidity tends to increase with age, and the analysis of multiple causes of death through death certificates in HICs has shown the number of causes reported increases with age (Désesquelles, 2015). This result appear as a good argument for the validity of this method.

Moreover, **women appear more suceptible to die from multiple causes of death than man accross age groups**. 2.4% of women's deaths compared to 2.0% of men's are identified as resulting from co-occurring causes, and this relationship holds true across age groups. This result appears consistent with the higher multimorbidity rates of women (Zhang et al., 2019), as they tend to suffer more from chronic diseases.

We can also see differences between HDSS sites (Appendix Table 9.3). However, these differences seem difficult to interpret considering the different age structure and epidemiological profile of their population, but also that, as the dataset was constructed reinterpreting retrospectively the VA questionnaires into 2012 questionnaires, they also might partly reflect differences in the questions present in the historical questionnaire in each site.

Table 5.2: Co-occurring causes by age group and sex

	All			**Female**			**Male**		
	% of all death	% of more than one cause	n	% of all death	% of more than one cause	n	% of all death	% of more than one cause	n
15-49 years	**1.3**	14.9	374	**1.5**	16.0	211	**1.2**	13.6	163
50-64 years	**2.3**	20.9	334	**2.6**	22.0	162	**2.1**	20.0	172
65 + years	**3.0**	24.4	883	**3.1**	25.6	475	**2.8**	23.1	408
All	**2.2**	20.6	1591	**2.4**	21.7	848	**2.0**	19.4	743

Reading : Among deceased aged from 15 to 49, across both sexes, 1.34% were identified as resulting from co-occurring causes, which represents 374 deaths and 14.86% of the deaths among deceased aged from 15 to 49 across both sexes that were attributed more than one cause by InterVA-4

From 72 330 VA of adults in 22 sites of the INDEPTH HDSS Network, 1992-2012 data

5.2 The importance of associations between NCDs and infectious diseases: a cumulative burden of disease at the individual level?

As illustrated in Table 5.3 (see Appendix Table 9.5 for the exhaustive table), a majority (61%) of the multiple causes identified consists of an association between non-communicable diseases and diseases of poverty, and a little less than third (30%) of associations between non-communicable diseases, with little multiple causes resulting from associations between diseases of poverty (6%), or involving injuries and violent deaths (less than 2%).

In the context of the health transition, the importance of associations between NCDs and diseases of poverty is particularly interesting to highlight. Those two categories of causes are often treated separately or even opposed in literature, as a result of the theory of health transition (Omran, 1971) on the one hand and differences in terms of public policy implementation. However, as other works on multiple causes of deaths have highlighted (Désesquelles, 2015), this dichotomy is much more nuanced on the individual level : non-communicable diseases are often risk factors for individual suffering from infectious diseases, as the recent Covid-19 pandemic could illustrate, associating a number of NCDs (such as hypertension, diabetes, chronic respiratory diseases...) with an increased mortality rate. Conversely, infectious diseases can create a fertile ground for the development of NCDs, this is for example the case for a number of sexually transmitted diseases increasing the risk of cancers.

Table 5.3: Co-occurring causes by group selected through the Euclidean index of similarity

Group A	Group B	Frequency	Percentages
Non-communicable diseases	**« Diseases of poverty »**	**976**	**61.3**
Diabetes and cardiovascular diseases	Infectious and parasitic diseases	648	40.7
Infectious and parasitic diseases	Other non-communicable diseases	189	11.9
Cancers	Infectious and parasitic diseases	60	3.8
Anemia and malnutrition	Diabetes and cardiovascular diseases	31	1.9
Chronic respiratory diseases	Infectious and parasitic diseases	25	1.6
Non-communicable diseases	**Non-communicable diseases**	**486**	**30.5**
Diabetes and cardiovascular diseases	Diabetes and cardiovascular diseases	165	10.4
Diabetes and cardiovascular diseases	Other non-communicable diseases	132	8.3
Cancers	Diabetes and cardiovascular diseases	111	7.0
Chronic respiratory diseases	Diabetes and cardiovascular diseases	39	2.5
Cancers	Other non-communicable diseases	28	1.8
« Diseases of poverty »	**« Diseases of poverty »**	**97**	**6.1**
Infectious and parasitic diseases	Infectious and parasitic diseases	69	4.3
Non-communicable diseases	**Injuries and violent deaths**	**29**	**1.8**

Associations irrespective of order attributed by InterVA-4, percentage >= 1.5%

1,591 VAs of adults with co-occurring causes of death (with a Euclidean index >= 0.65), INDEPTH Network, 1992-2013

This is especially important to underline in a context of a cumulative burden of disease quite different from the epidemiological profile of HICs, where burdens of infectious diseases remain significant while the burden of NCDs is increasing. **Our approach underlines the presence of a certain cumulative burden at the individual, with deaths resulting from a combination of infectious and non-communicable diseases, even though the proportion identified remains limited, representing a little more than 1 every 100 deaths**. The proportion of these associations among multiple causes identified can be considered to be artificially inflated, as we are not able to identify multiple causes with very similar symptoms, especially associations between infectious diseases. This can explain the relatively small portion of associations between diseases of poverty identified as co-occurring causes, and that are most probably importantly underestimated. However, even before the differentiation between co-occurring and competing causes, the proportion of associations between NCDs and diseases of poverty already represented around a third (31%) of deaths attributed more than one cause of death (for the whole table see Appendix, Table 9.6), suggesting its relative importance. This method seem to provide a first approach to analysing this cumulative burden of disease at an individual level, quite difficult to identify.

The association of chronic and acute diseases appear to be the most common, presenting an fruitful pattern to analyse multimorbidity. Table 5.4 illustrates the most frequent co-occurring causes identified, that consist in majority of an association between a chronic disease (mostly diabetes and cardiovascular diseases, as well as HIV/AIDS) with acute diseases such as strokes, acute cardiac diseases, acute respiratory diseases or acute abdomen (the exhaustive table of co-occurring causes identified available upon request) for the exhaustive list of associations identified). These co-occurrences underline that, beyond NCDs, a certain number infectious diseases also constitute important risk factors, such as HIV/AIDs representing close to 10% of the multiple causes identified. However, as we have seen, the importance of comorbidities with HIV/AIDs is very probably highly underestimated due to the definition of HIV in InterVA-4's matrix, making difficult to identify co-occurring causes with infectious diseases such as tuberculosis, that are in reality very common.

5.3 The important burden of cardiovascular diseases and diabetes in the multiple causes identified

Second, the importance of NCDs, especially diabetes and cardiovascular diseases among the multiple causes of death identified needs to be underlined. Indeed, more than two-thirds (69.8%) of the multiple causes identify result from an association with diabetes or cardiovascular diseases, whereas diabetes and cardiovascular diseases only represent 16.3% of the cause-specific mortality fraction of this population. In particular,

Table 5.4: Most frequent co-occurring causes identified

Cause A	Cause B	Frequency	Percentages
Diabetes mellitus	Stroke	148	9.3
Acute resp infect incl pneumonia	Other and unspecified cardiac dis	141	8.9
Pulmonary tuberculosis	Other and unspecified cardiac dis	123	7.7
Acute resp infect incl pneumonia	Acute cardiac disease	49	3.1
Stroke	Acute abdomen	48	3.0
Acute resp infect incl pneumonia	Stroke	44	2.8
Diabetes mellitus	HIV/AIDS related death	43	2.7
Other and unspecified cardiac dis	HIV/AIDS related death	43	2.7
Acute abdomen	HIV/AIDS related death	42	2.6
Stroke	Digestive neoplasms	42	2.6
Acute resp infect incl pneumonia	Acute abdomen	40	2.5
Stroke	Other and unspecified NCD	31	1.9

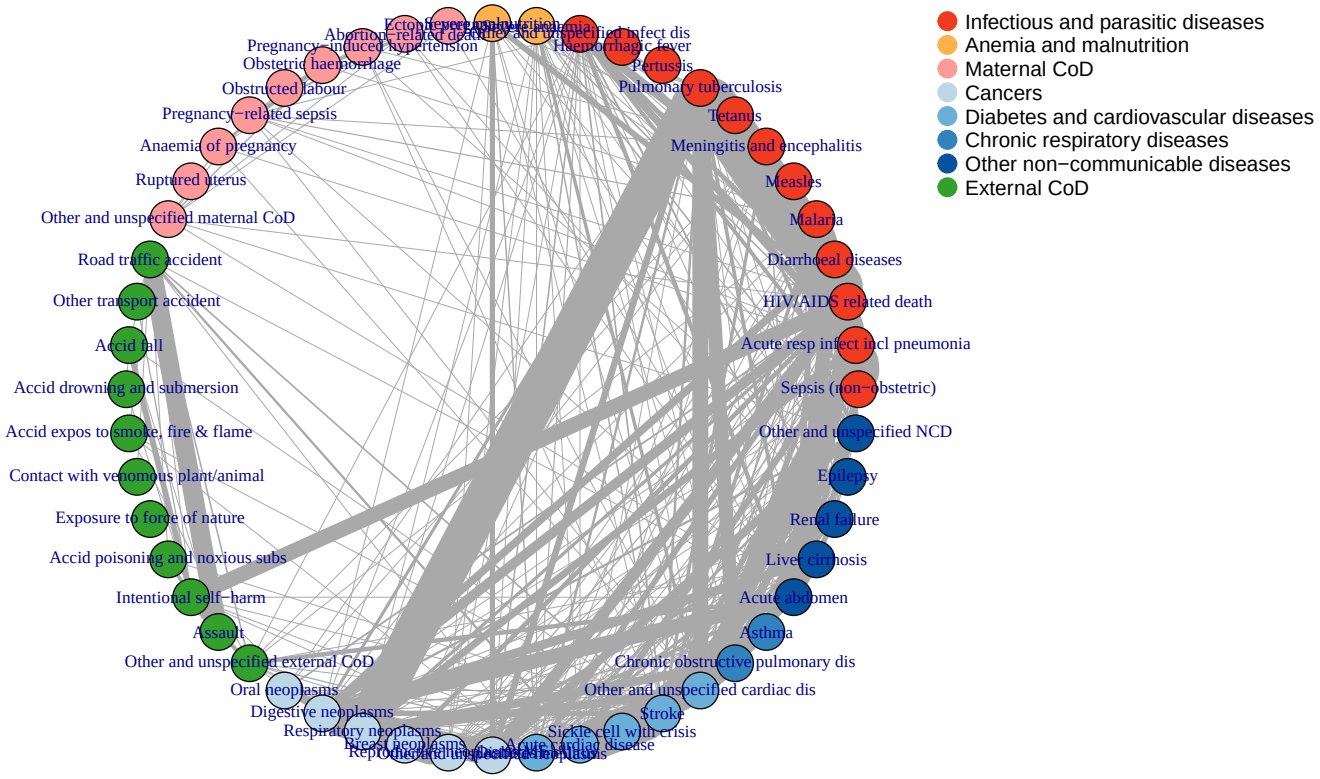
Associations irrespective of order attributed by InterVA-4, with frequency >=30

1,591 VAs of adults with co-occurring causes of death (with a Euclidean index >= 0.65), INDEPTH Network, 1992-2013

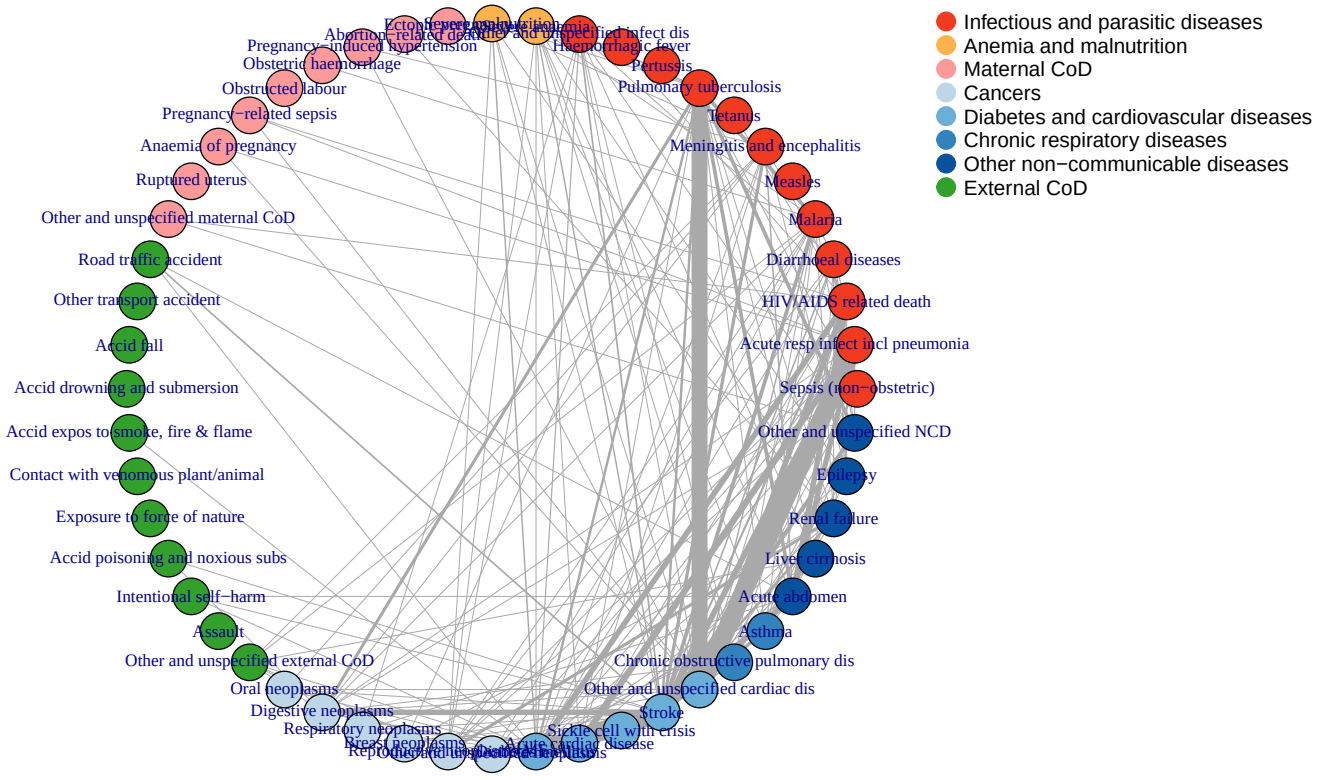
40.7% of multiple causes are associations between diabetes and cardiovascular diseases and infectious diseases. In a context were infectious diseases still represent the highest share among all causes of death amongst adults (see Table 2.2), this result appears as an the most striking illustration of the cumulative burden of disease at the individual level.

This importance is in part due to the associations of these chronic diseases with acute ones, but also to the importance of associations within NCDs, including cardiovascular diseases (representing 10% of all co-occurring causes). Indeed, at the most detailed level, the most frequent association of causes identified is diabetes mellitus and stroke (Table 5.4), representing 9% of co-occurring causes.

Figure 5.1: From more than one cause of death to co-occurring causes of deaths



(a) Associations of causes among the VAs attributed more than one cause of deaths by InterVA-4 (7,734 deaths, 10.7% of deaths)



(b) Co-occurring causes of death identified through the similarity index among the VAs attributed more than one cause of deaths (1,591 deaths, 2.2%)

6. Discussion : limitations and perspectives

We have exposed two different methods that present an important potential to identify multiple causes of death from verbal autopsies using InterVA-4. Both consist in identifying multiple causes amongst the deaths attributed more than one cause by the algorithm, and are based on its *a priori* probability matrix, but differ in the level of details required in the VAs data to operate.

- **The general approach provides an criterion to establish *a priori* a list of associations of causes to interpret as multiple causes and a list of associations to exclude as much as possible competing causes, using only the information available in the algorithm**. It has been the main focus of this paper,
- **The second method, the empirical approach, builds on this first method, and uses in addition the detailed information reported in the VA to evaluate for each death if a given association can be interpreted as multimorbidity**. Allowing for more precision, it however requires having access to the detailed information reported in VA questionnaires.

Underlining the strengths and limitations of these two complementary methods offers perspectives on multimorbidity monitoring through VAs but also on VA-methodology as a whole. We will first focus the hypotheses of the approach shared by both methods, to then compare the two methods, and finally conclude on the perspectives they offer on VAs as a methodology.

6.1 The difficulties of identifying multiple causes of death from VAs compared to death certificates

6.1.1 A conservative approach of multimorbidity...

One of the main characteristic of our methods and their hypotheses is that it builds a restrictive approach to multimorbidity, which explains the relatively low percentage of death (2,2%) identified as resulting from multiple causes.

First and foremost the approach of multimorbidity through multiple causes of death is in itself restrictive. It limits possible co-morbidities to diseases or situations considered causes of death, as defined by the ICD or by InterVA, excluding risk factors or conditions (such as hypertension, tobacco or alcohol consumption...) often taken into account in the study of multimorbidity (see section 1.4). This first restriction is a methodological choice : it provides a useful framework and a clear definition for the analysis of multimorbidity leading to death. Focusing on diseases rather than risk factors, and on mortality rather than morbidity, it is not directly comparable to most multimorbidity studies carried out in LMICs to this day, but offers a valuable complementary perspective.

However, our approach based on a similarity index is more conservative compared to a multiple causes of death approach based on death certificate because of the uncertain nature of the information provided by VAs, leading to probable underestimation. This is true for several reasons :

- **First and foremost, associations of causes with similar symptoms are not identifiable through this method**.
  This is probably the main limitation of this approach, making the identification of associations between most infectious diseases or between cancers impossible, in order to insure the exclusion of competing causes from this analysis. It seems very difficult to overcome, as it would require having access to the information of a positive test to both diseases or causes, an unrealistic expectation in a context where an important number of deaths occur away from health facilities[1]. Keeping in mind this limitation, this method guarantees to exclude as much

[1]The information about biological tests is only asked for malaria and HIV in the VA, and when reported, often leads to a unique diagnosis.

as possible competing causes ; it seems the safest strategy to provide estimations of associations of causes that can confidently be interpreted as multiple causes and constitute useful information about multimorbidity.

- **Moreover, the identification of multiple causes of death is inherently restricted by the lack of information present in VAs**, making the comparison with multiple causes identified through death certificates difficult. In many ways, the information collected from a the final caretaker of the deceased a few days to weeks after death risks to be incomplete compared to a diagnosis elaborated by a physician, a limitation that however underlies all VAs' analysis. This is why, algorithm such as InterVA were developed to analyse cause-specific mortality at the population level, but warn their users about the limitations of drawing conclusion at the individual level. With this approach, we divert InterVA from its original purpose to analyse cause-determination at the individual level; this is why we aim to put these uncertainty and limitations at the heart the approach.

- **Finally, the method depends on the strict and somewhat arbitrary selection rule of InterVA to attribute more than one cause of death**.
  Indeed, our method relies solely on the examination of the deceased attributed more than one cause of death by InterVA. As exposed in section 3.1.4, in order to be selected as the second likely cause of death, the associated probability of the second most probable cause - p_2 - has to be greater or equal to 1/2 of the likelihood of the most probable cause - p_1 - ($p_2 >= 1/2p_2$). This is a somewhat arbitrary rule, that can appear justified in order estimate cause-specific mortality fractions at the population level where we aim to identify underlying causes of death, however it appears quite restrictive when focusing on multimorbidity. Indeed, only 11% of deaths are attributed more than one cause according to this rule, with limited variations by sites and age group, when around a quarter (25%) of deaths all ages combined had more than one cause mentioned on their death certificate in France in 2011 (Désesquelles et al., 2016).
  It would be interesting to relax this rule, and allow for the considerations of more deaths with more than one cause, in order to widen the approach. This could be possible from the detailed data of VAs by using the detailed distribution of likelihoods across all the possible causes for each death. This also raises the question of taking into consideration more than the two most probable cause of death, especially if the criterion to select causes of death is relaxed. This would require some creativity on the methodological standpoint as it would necessitate considering triads of causes rather than diads. However, considering the number of death attributed three or more causes, this question remains marginal.

Considering all these limitations, the goal of this method is not to take our results at face value, but rather to consider them as useful estimations that can help monitor trends across time, space and demographic groups.

6.1.2 ... that does not automatically characterise the relationship between causes but provide an interesting tool to reflect on cause coding based on VAs

The second main limitation of this approach using algorithmic interpretations of VAs is that it does not qualify the relationship between causes. **Indeed, compared to death certificates, the probabilistic approach used by InterVA puts all causes on the same level and does not describe the relationship between them**. On the contrary, causes of death in death certificates are coded in order to determine which is the underlying cause, which can be considered as the immediate cause or as the intermediate cause. This information would be very valuable, as it allows to understand the sequencing of the causes, and set appropriate priorities for public health policies.

However, characterising the relationship between causes is a complex process. For death certificates alone, the coding process relies on an important series of norms and rules defined by the WHO as part of the ICD that have been constructed since the XIXth century by the international community (Star and Bowker, 1999). Considering this complexity, the automatic characterisation of the relationship between causes based on VAs seems a very ambitious goal, especially as the consideration of multiple causes of death has not been at the center of the elaboration of VAs as a tool.

However, our approach allows for a case-by-case appreciation of the possible relationship between causes. Moreover, it provides for an interesting thoughts about further developments of for cause coding from VAs, as it aims to take better into account the question of multiple causes of death and the strategies

possible for their coding process. These consideration seems to be on the agenda of WHO VA Reference Group now part of the WHO Family of International Classifications (WHO-FIC) Network. These analyses could contribute to their reflections, while this methodology would in turn benefit from implementations made in this direction.

6.2 Using a index of similarity between causes : from the general approach to the empirical approach

6.2.1 The theoretical approach : a pragmatic and robust approach when detailed VAs are not available ...

Our first theoretical approach appears as a pragmatic method to estimate multiple causes of death when detailed VAs are not available. It provide a rather simple and useful tool to differentiate between competing and co-occurring causes.

As all approaches relying on an index, it remains dependent on the specific index formula and the threshold chosen. This is why we decided to test different formulas to compute distance between causes, and also tested different thresholds. This test have shown the methods appears quite robust, with very similar results. More elaborated index of similarity formulas could also be tested ; however, considering the already limited nature of VAs, it seemed more reasonable to stick the simpler functions, such as norms, to evaluate distance.

This consideration of robustness also rises the question of **the dependence of our index to the function that transforms the probability matrix in letters (that is set by the experts) to numbers (probabilities used by the model)**. Indeed, the transformation function used by InterVA corresponds currently to a logarithmic rather than a linear perception of probabilities (see section 3.1.3 Table 3.1), which seems to correspond to the human perception of qualitative expressions of probability in medical settings (Ohnishi, 2002). However, this transformation function can appear quite arbitrary, and questions remain about the effects of changing this function on the model. However, this question is however a little out of scope of our current study ; we aim here to build on the current model, taking its diagnosis and in this transformation function, as a given. Nevertheless, the difference between the Euclidean index and the absolute norm index could contribute to answer this question. Indeed, by putting less weight on greater differences, the absolute norm function is reducing the numerical differences between the different grades attributed by the experts, while the Euclidean index is making this difference greater. The important similarity between the two indexes suggest that on this scale the differences remain small, however, testing with more different transformation functions could be made.

6.2.2 ... that could be refined by considering some characteristics of the deceased (age group and sex)...

This theoretical approach remains quite a blunt tool, as no individual characteristics (geographic origin, age groups and sex) are taken into account to differentiate between competing and co-occurring causes. The empirical approach on the other hand allows to take into account all the reported characteristics of the deceased, but requires access to detailed VAs. However, **the theoretical approach could be refined to take into account more individual characteristics that are available (sex, age group and site) without access to detailed VAs**. It is what we aim to discuss here :

- First, **we could adapt the index to the epidemiological profile of each site (HIV and malaria prevalence)**, as these prevalence are asked by InterVA to make diagnoses and changes the *a prior* matrix[2]. However, these changes are very marginal when taking into account all 246 indicators to compute the index, and do not change the assessment of multiple causes. Comparing the two extreme settings, High High versus Very Low Very Low, Table 6.1 illustrate the very minimal difference : the maximum difference between the indexes does not exceed 0.013, with only 10 associations where the difference is greater than 0.008 (for details,

[2]As explained in Section 3.1.3, the prevalence of malaria and HIV can be separately set to High (0.05), Low (0.005) or Very Low (0.00001)).

see Appendix, Table 9.7), and the index remains the same for more than half of associations (from the first quartile to the 3rd quartile) when considering five significant digits.

Therefore, **taking into account these difference when computing an index from the 246 indicators appears unnecessary**, this is why we decided to use the default *a priori* matrix (HIV = very low, malaria disease = very low) for this paper - even though, there are no site in the INDEPTH database declaring these prevalences (for the detail of deaths by prevalence see Appendix, Table 9.8). **However, if we refine the indicators taken into account, it can become an important factor to consider**; this is especially true for the empirical index, where it seems crucial to use the probability matrix with the appropriate settings to compute the index.

Table 6.1: Distribution of the difference between the theoretical Euclidean indexes based on the a priori matrix with prevalence of malaria and HIV set both to very low (VLVL) compared to high (HH)

	Difference VLVL - HH
Min.	-0.00199
1st Qu.	0.00000
Median	0.00000
Mean	0.00020
3rd Qu.	0.00000
Max.	0.01312

From the a priori matrices of InterVA-4

- On the other hand, **taking into account the sex and age group of the deceased would be an interesting way of refining the index the index of similarity**, as the information is available to compute cause-specific mortality fractions, even when detailed VAs are not accessible. Instead of computing an index from the entire *a priori* matrix, indexes specific to each sex and/or age group could be calculated, by filtering only questions relevant to each category. This would mean for example that indicators regarding pregnancy would not be considered for males, and questions specific to new borns would not be considered for adults, reducing the 246 indicators to a little more than 60 for each category. This could constitute as a middle ground between the general theoretical index and the empirical. But the impact of such a refinement on results still needs to be evaluated.

6.2.3 ... that remains limited by the heterogeneity of VAs and the INDEPTH database

- **On a local level, the results presented above our limited by the heterogeneity across sites of the INDEPTH database**. Indeed, the database used is the result of a considerable work to be able to compare the mortality profiles of as much sites of the INDEPTH Network (Streatfield, 2014, (1), (2) and (3)), that was carried out before the important standardisation of VAs through InterVA (with InterVA-4 released in 2012, while the data covers a period from 1992 to 2012). Hence, the database has been constructed by retrospectively recoding the VA questionnaires of each sites into the 2012 WHO questionnaire used by InterVA-4, to then determine causes of death. This means that heterogeneity between sites could be important, based on the difference between the questionnaire of each site and the InterVA-4 questionnaire, even though the extent of these differences is not entirely clear but occasionally mentioned the articles of the creators of the database (Streatfield (1), (4)).
  This is in part why it appeared difficult to comment the differences of multiple causes between sites, as the margin of error of the index might differ between sites if some questions did not exist in certain sites. **However, this remains a local limitation that future analysis will very likely overcome : since 2012, the standardisation of VA questionnaires has considerably progressed, through the WHO efforts and the spread of the use of InterVA**.

- **More generally, beyond consideration of sex and age groups, the theoretical approach is limited by the heterogeneity of information reported in VAs**. Indeed, as the test of robustness by the empirical index has shown (Section 4.4.2 Figure 4.6), depending on the symptoms reported, the associations of two causes can appear as competing or co-occurring causes. This is especially true for causes that *a priori* seem very different, but could be confused if very little symptoms are reported. Hence, while this theoretical method remains very cautious, it appears that some associations considered *a priori* co-occurring causes, can prove to be competing with a closer inspection of the symptoms reported, even though this case remains limited. It is this main limitation that leads us to the empirical approach as more precise evaluation tool, when detailed VAs are available.

6.2.4 The empirical approach : a more precise evaluation, that could be completed by analysing reported chronic or pre-existing conditions, and that raises the question of an approach in terms of clusters of symptoms

The empirical index of similarity seems to answer to most of these limitation, when access to the detailed VAs is possible. Building on the strength of the theoretical approach, it allows for more precision as it bases the index exactly on the information used by the algorithm to determine the cause of death. It seems a promising approach, that however requires testing on more VAs data to provide firmer conclusions that could help nuance and interpret the theoretical approach as well.

The main limitation of this approach is dependent on a limitation of InterVA's model. Whereas all questions of VA questionnaires give the possibility to answer: "Yes", "No", "I don't know", the algorithm doesn't differentiate between "No" and "I don't know", only taking into account positive answers. However, "No" answers could be considered valuable information, provided that the interviewee does recognise correctly the limitations of his knowledge on the deceased symptoms. Other algorithms, such as InSilicoVA (Clark et al., 2015), are working on taking into account these differences in the process of cause determination. It could be an interesting track to adapt our approach to this other algorithm also based on the same *a priori* matrix.

Moreover, having access to detailed VAs raises further question about possible refinement of the method. Indeed, **rather than looking at symptoms independently, as is done by InterVA, it could be interesting to aim to identify clusters of symptoms**. This would possibly allow for a more precise and accurate identification of multimorbidity (though a much more intricate methodology), and appears as a very interesting path forward to investigate multimorbidity through VAs. However, it remains out of scope of the current study, as it would require very precise and good quality data about symptom patterns and an entire new methodology. Moreover, the intricacy of the methodology involved should be weight against the precision offered by VA as a tool. Indeed, as reminded by Fottrell et al (Fottrell et al, 2011), *"by keeping in mind who needs cause of death data and for what purposes, reasonable degrees of imprecision become acceptable and the criteria of efficiency, affordability and reliability become paramount"*. Considering the imprecision of the data, the use of very intricate methodology could appear counter-productive, possibly concealing the limitation of the data, in particular the important amount of symptoms that are not reported.

On a more general note, **it seems that simply taking into account the pre-existing conditions, chronic diseases and risk factors reported in VAs could importantly benefit the analysis of multiple causes and multimorbidity**. Indeed, the 2012 VA questionnaire asks 17 questions about the medical history of the deceased, providing information about a possible diagnosis of : heart disease, tuberculosis, HIV/AIDS, high blood pressure, diabetes, asthma, epilepsy, cancer, chronic obstructive pulmonary disease (COPD), dementia, depression, stroke, sickle cell disease, kidney disease, liver disease, measles (and an extra question on a possible recent positive or negative test to malaria). These questions contribute to the interpretation of causes of death, however, they do not determine it. An important part of the pre-existing conditions do not appear as causes of death, as illustrated by Table 6.2. Indeed, more than half (58%) of deceased that were reported as having a history of HIV/AIDs by their final caretaker had no mention of HIV/AIDs in the causes of death selected by InterVA, and a quarter (24%) of deceased with a reported history of cancer had no mention of cancer among their causes of death. This proportion is even more important for individual suffering from diabetes, as 85% that had declared a history of diabetes had no mention of diabetes among their causes of death.

It shows that nformation of possible chronic or pre-existing condition do not necessarily determine the underlying cause whereas we could assume that, by default, such condition could be consider as multimorbidity. This might constitute the main limitation of the empirical index approach, solely based on the results of the algorithm. On the other hand, analysing this information about the medical history of the deceased in relation to the causes of death determined by the model could be valuable. Moreover, it would allow to open the lens from multiple causes of death, to a broader conception of multimorbidity taking into account risk factors such as hypertension, but also alcohol and tobacco consumption, that are also part of the information collected during VAs. This would require a selection of the medical history question to consider, as they do not seem to offer the same quality of information[3], but would offer a simple way of incorporating multimorbidity monitoring into VAs.

Table 6.2: Reported conditions and causes of death determined by InterVA-4, VAs from the HDSS of Ouagadougou Burkina Faso

(a) History of HIV/AIDs and causes of death

	HIV among causes of death	No HIV among causes of death	n	% total
No reported history of HIV/AIDs	4.2	95.8	1690	98.6
Reported history of HIV/AIDs	41.7	58.3	24	1.4
Ensemble	4.7	95.3	1714	100.0

(b) History of diabetes and causes of death

	Diabetes among causes of death	No diabetes among causes of death	n	% total
No reported history of diabetes	1.0	99.0	1635	95.4
Reported history of diabetes	15.2	84.8	79	4.6
Ensemble	1.7	98.3	1714	100.0

(c) History of cancer and causes of death

	Cancer among causes of death	No Cancer among causes of death	n	% total
No reported history of cancer	5.3	94.7	1630	95.1
Reported history of cancer	76.2	23.8	84	4.9
Ensemble	8.8	91.2	1714	100.0

From 1714 VAs of all age groups, collected between 2010 and 2019, HDSS of Ouagadougou Burkina Faso
***Reading :* 24 deceased were reported as having a history of HIV/AIDs, among them, 41.7% had HIV/AIDs related death among the causes determined by InterVA-4, and 58.3% had no mention of HIV/AIDs among the causes selected by InterVA.**

7. Conclusion : new perspectives on VA-methodology

7.1 A reflection on the limitations of VAs and the definition of each cause of death by the *a priori* probability matrix

All in all, with this report, we aim to demonstrate the potentiality VAs to routinely monitor multiple causes of death from probabilistic algorithms of interpretation. Focusing on InterVA-4, we provide two methods to identify multiple causes and first estimates of multiple causes mortality statistics on the scale of 22 HDSS sites across Asia and Sub-saharan Africa, in order to proove the concept.

[3]For instance depression and dementia are not identified in survey to general population with a simple direct question like in VAs, these informations in VA can be considered limited.

First of all, this reflection provides an interesting perspective on VA methodology as a whole. By focusing on possible confusions between causes, the similarity index offers an interesting framework to apprehend the definition of causes by the *a priori* matrix and their possible shortcomings. It underlines the difficulty of defining causes such as HIV/AIDs or chronic diseases that often indirectly cause death through an opportunistic disease, that can be partially taken into account by the definition but cannot insure the coding of the chronic condition as the underlying cause of death depending on the number of other symptoms declared. Moreover, underlining the difference between the theoretical and empirical indexes appears especially valuable, as it compares the set of information we would ideally access with the information actually collected, limited by the conditions of a VA interview, a difficult and sometimes uncomfortable task for both the interviewer and the grieving caretakers, constrained by the lack of knowledge and health care related to adults health issues in particular.

7.2 An important potentiality for routinely monitoring multimorbidity leading to death, adaptable to a wide range of algorithms

All in all, while some refinement could be considered, and the method remains dependent on the limited information provided by VAs, it seems important to underline the adaptability of this approach to other probabilistic algorithms existing and still in development, as the research in VA interpretation progresses. Indeed, focused on InterVA-4, this method can be directly adapted to all other version of InterVA, especially the latest version InterVA-5, here only the number of indicators and some part of the *a priori* matrix have been changed. Adaptations to other algorithms such as InSilicoVA or other future models aiming to determine causes of death from VA without "gold standard" data appears also particularly interesting, as they will all likely rely on an probability matrix. InSilicoVA (Clark et al., 2015) particularly builds on InterVA's model and its probability matrix to elaborate a more complex model, and aims to address a certain number of limitations of InterVA discussed above : among others its transformation function from ranking to *a priori* probabilities of symptoms associated to each cause, and its consideration of only reported symptoms and not the absence of symptoms. In particular, InSilicoVA aims to differentiate between the absence of a symptom and an unanswered or unknown answer to a question, that seems promissing.

Automated models of interpretation of VAs appear as the way forward to developing intermediate CRVS systems in contexts constrained by resources; indeed, not only VA is a methodology that is spreading and standardising through HDSS site development, it is also starting to develop outside HDSS as potential tools to establish national cause-specific statistics through phone based interviews, for samples of the population or the whole population. Recent research from HICs has shown the importance of taking into account multimorbidity and especially multiple causes of death to understand the morbid processes leading to death, a realisation only emphasised by the recent Covid-19 pandemic, where the important role of co-morbidities as risk factors has been underlined. The relevance of this framework seems only heightened in the context of LMICs, undergoing a health transition notably marked by a cumulative burden of infectious an non-communicable diseases, and that appear a priority in understanding the impact of a pandemic such as Covid-19 as few recent results show (Kirenga and Pauline Byakika-Kibwika, 2021). However, it remains largely unmonitored ; **this approach aims to provide perspective to envisage incorporating the monitoring multiple causes of death in future models, and feed the reflection around this possibility**.

One of the key challenge in the monitoring multiple causes and co-morbidities from VAs in the long run will be the integration of the WHO cause-coding rules, allowing to qualify causes as underlying or associated. This project aims to advocate for this integration, and provide some thoughts for its elaboration.

8. Bibliography

Academy of Medical Sciences. 2018. Multimorbidity: A priority for global health research. London: Academy of Medical Sciences, 2018. https://acmedsci.ac.uk/policy/policy-projects/multimorbidity.

Almirall, José, and Fortin, Martin . 2013. « The Coexistence of Terms to Describe the Presence of Multiple Concurrent Diseases »: Journal of Comorbidity. https://journals.sagepub.com/doi/10.15256/joc.2013.3.22 (1 décembre 2020)

Banerjee, Amitava, John Hurst, Edward Fottrell, et J. Jaime Miranda. 2020. « Multimorbidity: Not Just for the West ». Global Heart 15(1): 45.

Boutayeb, Abdesslam. 2006. « The double burden of communicable and non-communicable diseases in developing countries ». Transactions of The Royal Society of Tropical Medicine and Hygiene 100(3): 191-99.

Byass Peter, Chandramohan Daniel, Clark Samuel J., Lucia D'Ambruoso,Edward Fottrell, Wendy J. Graham, Abraham J. Herbst, Abraham Hodgson, Sennen Hounton,Kathleen Kahn, Anand Krishnan, Jordana Leitao, Frank Odhiambo, Osman A. Sankoh and Stephen M. Tollman , 2012. « Strengthening standardised interpretation of verbal autopsy data: the new InterVA-4 tool », Global Health Action, 5:1, 19281, DOI:

Clark, Samuel J., Tyler McCormick, Zehang Li, et Jon Wakefield. 2015. « InSilicoVA: A Method to Automate Cause of Death Assignment for Verbal Autopsy ». arXiv:1504.02129 [stat]. http://arxiv.org/abs/1504.02129 (2 juin 2021).

Delaunay, Valérie. 2018. « L'apport spécifique des Observatoires de Population à la connaissance de l'évolution des sociétés au Sud : Example du champ des études sur l'enfance » : II.1. Les observatoire de population. Mémoire pour l'obtention de l'habilitation à diriger des recherches.
‘

Désesquelles, Aline et al. 2010. « Revisiting the Mortality of France and Italy with the Multiple-Cause-of-Death Approach ». Demographic Research 23: 771-806.

Désesquelles, Aline et al. 2015. « After the Epidemiologic Transition: A Reassessment of Mortality from Infectious Diseases among over-65s in France and Italy ». International Journal of Public Health 60(8): 961-67.

Désesquelles Aline, Gamboni Andrea, Demuru Elena, et the MultiCause Network. 2016. « We Only Die Once... but from How Many Causes? » Population Societies No 534(6): 1-4. https://www.cairn-int.info/revue-population-and-societies-2016-6-page-1.html

Duthé, Géraldine, Rossier, Clémentine, and Bassiahi Soura, Abdramane. 2019. *Inégalités de santé à Ouagadougou : Résultats d'un observatoire de population urbaine au Burkina Faso*. Paris, Ined Éditions.

Fottrell, Edward, Kathleen Kahn, Stephen Tollman, et Peter Byass. 2011. « Probabilistic Methods for Verbal Autopsy Interpretation: InterVA Robustness in Relation to Variations in A Priori Probabilities ». PLOS ONE 6(11): e27200.

Ford, Joanna C., and Ford, John A. . 2018. « Multimorbidity: Will It Stand the Test of Time? » Age and Ageing 47(1): 6-8.

Fortin, Martin et al. 2012. « A Systematic Review of Prevalence Studies on Multimorbidity: Toward a More Uniform Methodology ». The Annals of Family Medicine 10(2): 142-51.

Fuhrman, Claire. 2014. « Surveillance épidéimiologique de la multimorbidité : Revue bibliographique ». Institut de veille statistique, Département des maladies chroniques et des traumastismes, Saint Maurice.

Garenne, Michel, et Vincent Faveau. 2006. « Potential and Limits of Verbal Autopsies ». Bulletin of the World Health Organisation 84(3): 164-164.

INDEPTH Network. INDEPTH Network Cause-Specific Mortality - Release 2014. Oct 2014. Provided by the INDEPTH Network Data Repository.

Johnston, Marjorie C. and al. 2019. « Defining and Measuring Multimorbidity: A Systematic Review of Systematic Reviews ». European Journal of Public Health 29(1): 182-89.

Källander, Karin, Jesca Nsungwa-Sabiiti, et Stefan Peterson. 2004. «Symptom Overlap for Malaria and Pneumonia—Policy Implications for Home Management Strategies». Acta Tropica 90(2): 211-14.

Kirenga, Bruce J., et Pauline Byakika-Kibwika. 2021. « Excess COVID-19 Mortality among Critically Ill Patients in Africa ». The Lancet 397(10288): 1860-61.

Kolčić, Ivana. 2012. « Double burden of malnutrition: A silent driver of double burden of disease in low– and middle–income countries ». Journal of Global Health 2(2). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3529312/ (11 février 2021).

Mathers, Colin D. 2020. « History of global burden of disease assessment at the World Health Organization ». Archives of Public Health 78(1): 77.

Martini, Jessica, et Audrey Figg. 2010. « 6. L'émergence du diabète de type 2 en tant que problème de santé publique ». In Santé internationale: les enjeux de santé au sud, Paris: Presses de Sciences Po, 105 à 118. https://www.cairn.info/sante-internationale---page-105.html

McCormick, Tyler H. et al. 2015. « Probabilistic Cause-of-Death Assignment Using Verbal Autopsies ». arXiv:1411.3042 [stat]. http://arxiv.org/abs/1411.3042 (24 septembre 2020).

Meslé, France, and Vallin, Jacques. 1988. « Chapitre 1 : Les origines » In.*Les causes de décès en France de 1925 à 1978*. Travaux et Documents, Cahier numéro 115, Institut National d'Etude Démographique, Presses Universitaires de France.

Meslé, France, and Vallin, Jacques. 2002. « La transition sanitaire : tendances et perspectives ». In *Démographie : analyse et synthèse*. III Les déterminants de la mortalité, p..

Mikkelsen, Lene et al. 2015. « A Global Assessment of Civil Registration and Vital Statistics Systems: Monitoring Data Quality and Progress ». The Lancet 386(10001): 1395-1406.

Ohnishi, M. 2002. « Interpretation of and Preference for Probability Expressions among Japanese Patients and Physicians ». Family Practice 19(1): 7-11.

Olshansky, S. J., et A. B. Ault. 1986. « The Fourth Stage of the Epidemiologic Transition: The Age of Delayed Degenerative Diseases ». *The Milbank Quarterly* 64(3): 355-91.

Omran, A.R., 1971, « The epidemiologic transition: a theory of the epidemiology of population change. 1971 », *The Milbank Quarterly*, 83(4), p/ 731-757.

Pati et al., 2015 « Prevalence and Outcomes of Multimorbidity in South Asia: A Systematic Review ». BMJ Open 5(10): e007235.

Prados-Torres, Alexandra et al. 2014. « Multimorbidity Patterns: A Systematic Review ». Journal of Clinical Epidemiology 67(3): 254-66.

Rao, Chalapati, et Mamta Kansal. 2020. « India's Civil Registration System: A Potentially Viable Data Source for Reliable Subnational Mortality Measurement ». medRxiv: 2020.04.03.20052894.

Remais, Justin V et al. 2013. « Convergence of non-communicable and infectious diseases in low- and middle-income countries ». International Journal of Epidemiology 42(1): 221-27.

Sankoh, Osman et on behalf of the INDEPTH Network and partners. 2017. « Why Population-Based Data Are Crucial to Achieving the Sustainable Development Goals ». International Journal of Epidemiology 46(1): 4-7.

Setel, Philip W. et al. 2007. « A Scandal of Invisibility: Making Everyone Count by Counting Everyone ». The Lancet 370(9598): 1569-77.

Star, Susan Leigh, et Geoffrey G. Bowker. 1999. Sorting Things Out: Classification and Its Consequences. MIT Press.

Streatfield, P. Kim et al. 2014. (1) « Cause-specific childhood mortality in Africa and Asia: evidence from INDEPTH health and demographic surveillance system sites ». Global Health Action 7(1): 25363.

Streatfield, P. Kim et al. 2014. (2) « HIV/AIDS-Related Mortality in Africa and Asia: Evidence from INDEPTH Health and Demographic Surveillance System Sites ». Global Health Action 7(1): 25370.

Streatfield, P. Kim et al. 2014. (3) « Pregnancy-related mortality in Africa and Asia: evidence from INDEPTH Health and Demographic Surveillance System sites ». Global Health Action 7(1): 25368.

Streatfield, P. Kim et al. 2014. (4) « Adult non-communicable disease mortality in Africa and Asia: evidence from INDEPTH Health and Demographic Surveillance System sites ». Global Health Action 7(1): 25365.

Thomas, Lisa-Marie, Lucia D'Ambruoso, et Dina Balabanova. 2018. « Verbal Autopsy in Health Policy and Systems: A Literature Review ». BMJ Global Health 3(2): e000639.

Tichenor M and Sridhar D. 2020. Metric partnerships: global burden of disease estimates within the World Bank, the World Health Organisation and the Institute for Health Metrics and Evaluation [version 2; peer review: 2 approved, 1 approved with reservations]. Wellcome Open Res 2020, 4:35 (https://doi.org/10.12688/wellcomeopenres.15011.2)

Willadsen, Tora Grauers and al. 2016. « The role of diseases, risk factors and symptoms in the definition of multimorbidity – a systematic review ». Scandinavian Journal of Primary Health Care 34(2): 112-21.

9. Appendix

9.1 Dictionnary of acronyms and abbreviations

COPD: Chronic obstructive pulmonary disease.

CRVS: Civil registration and vital statistics.

HDSS: Health and Demographic Surveillance System.

HICs: High Income countries.

INDEPTH Network: The International Network for the Demographic Evaluation of Populations and Their Health, for more information see section 2.1.

InterVA: a software designed to interpret verbal autopsies, i.e. determine probable causes of death. For more information and details about mechanisms used by the software see 3.1.

LMICs: Low and Middle income countries.

NCDs: Non-communicable diseases.

VA: Verbal autopsies.

WHO: World Health Organisation.

9.2 Tables and Figures

Figure 9.1: The Euclidean index heatmap of the associations present in the INDEPTH data

Table 9.1: All possible causes of death as defined by InterVA-4

Sepsis (non-obstetric)
Acute resp infect incl pneumonia
HIV/AIDS related death
Diarrhoeal diseases
Malaria
Measles
Meningitis and encephalitis
Tetanus
Pulmonary tuberculosis
Pertussis
Haemorrhagic fever
Other and unspecified infect dis
Oral neoplasms
Digestive neoplasms
Respiratory neoplasms
Breast neoplasms
Reproductive neoplasms MF
Other and unspecified neoplasms
Severe anaemia
Severe malnutrition
Diabetes mellitus
Acute cardiac disease
Sickle cell with crisis
Stroke
Other and unspecified cardiac dis
Chronic obstructive pulmonary dis
Asthma
Acute abdomen
Liver cirrhosis
Renal failure
Epilepsy
Other and unspecified NCD
Congenital malformation
Prematurity
Birth asphyxia
Neonatal pneumonia
Neonatal sepsis
Other and unspecified neonatal CoD
Fresh stillbirth
Macerated stillbirth
Road traffic accident
Other transport accident
Accid fall
Accid drowning and submersion
Accid expos to smoke, fire & flame
Contact with venomous plant/animal
Exposure to force of nature
Accid poisoning and noxious subs
Intentional self-harm
Assault
Other and unspecified external CoD
Ectopic pregnancy
Abortion-related death
Pregnancy-induced hypertension
Obstetric haemorrhage
Obstructed labour
Pregnancy-related sepsis
Anaemia of pregnancy
Ruptured uterus
Other and unspecified maternal CoD
Cause of death unknown

Table 9.2: Percentage of deaths with more than one cause attributed by InterVA-4, INDEPTH

	Frequency	Percentages
By sites		
Burkina Faso, Nouna	318	10.3
Burkina Faso, Ouagadougou	51	11.5
Côte d'Ivoire, Taabo	51	13.6
Ethiopia, Kilite Awlaelo	32	9.8
Ghana, Dodowa	290	11.4
Ghana, Navrongo	1201	14.4
The Gambia, Farafenni	191	11.9
India, Ballabgarh	200	11.9
India, Vadu	70	13.0
Indonesia, Purworejo	77	9.9
Kenya, Kilifi	331	10.0
Kenya, Kisumu	1398	11.7
Kenya, Nairobi	266	11.2
Malawi, Karonga	127	9.3
Senegal, Bandafassi	165	15.9
South Africa, Africa Centre	509	5.6
South Africa, Agincourt	994	10.8
Vietnam, Filabavi	77	11.1
Bangladesh, AMK	254	9.6
Bangladesh, Bandarban	28	12.1
Bangladesh, Chakaria	88	9.6
Bangladesh, Matlab	1016	10.4
By demographic characteristics		
Sex		
female	3908	11.1
male	3826	10.3
Age group		
15-49 years	2515	9.0
50-64 years	1598	11.1
65 + years	3621	12.1

From 72 330 VA of adults in 22 sites of the INDEPTH HDSS Network, 1992-2012 data

Figure 9.2: Most important indicators associated with HIV/AIDs related deaths according to InterVA-4

HIV/AIDs related death : symptoms with a probability >= 0.2 according to InterVA-4

B_HIVAIDS.C.7	names
0.8	Duration of final illness 3 weeks or more
0.8	History of HIV/AIDS
0.8	Not pregnant within 6 weeks of death
0.5	Age 15-49 years
0.5	Male
0.5	Female
0.5	Wet season
0.5	Dry season
0.5	Cough of any kind
0.5	Diarrhoea lasting 4 weeks or more
0.5	Headache
0.5	Shingles/herpes zoster
0.5	Became very thin or wasted
0.5	Received vaccines as appropriate for age at death
0.5	Received treatment for the illness that lead to death
0.5	Discharged from hospital very ill
0.2	Age 50-64 years
0.2	Fever lasting 2 weeks or more
0.2	Cough lasting 3 weeks or more
0.2	Diarrhoea of any kind
0.2	Vomiting
0.2	Weight loss
0.2	Sores or white patches in the mouth or tongue
0.2	Lumps/swelling in the neck

Figure 9.3: Most important indicators associated with Diabetes mellitus according to InterVA-4

Diabetes mellitus : symptoms with a probability >= 0.5 according to InterVA-4

B_DIABET.C.7	names
1	History of diabetes
0.8	Ulcers, abscess, sores on feet
0.8	Unconscious for at least 24 hours before death
0.8	Increased frequency of urination
0.8	Excessive water intake
0.5	Age 65+ years
0.5	Male
0.5	Female
0.5	Duration of final illness 3 weeks or more
0.5	Dry season
0.5	Any skin problems
0.5	Received vaccines as appropriate for age at death
0.5	Received treatment for the illness that lead to death

Figure 9.4: Most important indicators associated with Stroke according to InterVA-4

Stroke : symptoms with a probability >= 0.5 according to InterVA-4

B_STROKE.C.7	names
1	History of stroke
1	Paralysis of one side of the body
0.8	Died suddenly
0.8	History of high blood pressure
0.5	Age 65+ years
0.5	Male
0.5	Female
0.5	Duration of final illness < 3 weeks
0.5	Dry season
0.5	Headache
0.5	Unconsciousness started suddenly
0.5	Difficulty or pain while swallowing liquids
0.5	Received vaccines as appropriate for age at death
0.5	Received (or needed) treatment/food through nose

Table 9.4: Co-occurring causes by group selected through the Euclidean index of similarity (exhaustive)

Group A	Group B	Frequency	Percentages
Non-communicable diseases	**« Diseases of poverty »**	**976**	**61.3**
Diabetes and cardiovascular diseases	Infectious and parasitic diseases	648	40.7
Infectious and parasitic diseases	Other non-communicable diseases	189	11.9
Cancers	Infectious and parasitic diseases	60	3.8
Anemia and malnutrition	Diabetes and cardiovascular diseases	31	1.9
Chronic respiratory diseases	Infectious and parasitic diseases	25	1.6
Anemia and malnutrition	Chronic respiratory diseases	9	0.6
Anemia and malnutrition	Cancers	6	0.4
Anemia and malnutrition	Other non-communicable diseases	3	0.2
Diabetes and cardiovascular diseases	Maternal CoD	3	0.2
Other non-communicable diseases	Maternal CoD	2	0.1
Non-communicable diseases	**Non-communicable diseases**	**486**	**30.5**
Cancers	Diabetes and cardiovascular diseases	111	7.0
Diabetes and cardiovascular diseases	Other non-communicable diseases	132	8.3
Cancers	Chronic respiratory diseases	3	0.2
Diabetes and cardiovascular diseases	Diabetes and cardiovascular diseases	165	10.4
Cancers	Other non-communicable diseases	28	1.8
Chronic respiratory diseases	Other non-communicable diseases	7	0.4
Other non-communicable diseases	Other non-communicable diseases	1	0.1
Infectious and parasitic diseases	Maternal CoD	13	0.8
« Diseases of poverty »	**« Diseases of poverty »**	**97**	**6.1**
Infectious and parasitic diseases	Infectious and parasitic diseases	69	4.3
Anemia and malnutrition	Infectious and parasitic diseases	13	0.8
Anemia and malnutrition	Anemia and malnutrition	2	0.1
Non-communicable diseases	**Injuries and violent deaths**	**29**	**1.8**
Diabetes and cardiovascular diseases	External CoD	21	1.3
Cancers	External CoD	4	0.3
Other non-communicable diseases	External CoD	4	0.3
« Diseases of poverty »	**Injuries and violent deaths**	**3**	**0.2**
Infectious and parasitic diseases	External CoD	3	0.2

Associations irrespective of order attributed by InterVA-4

1,591 VAs of adults with co-occurring causes of death (a Euclidean index >= 0.65), INDEPTH Network, 1992-2013

Table 9.3: Co-occurring causes by HDSS sites

	% of all death	% of more than one cause	n
Côte d'Ivoire, Taabo	2.67	19.61	10
The Gambia, Farafenni	2.31	19.37	37
Bangladesh, AMK	2.16	22.44	57
Bangladesh, Bandarban	1.73	14.29	4
Bangladesh, Chakaria	2.40	25.00	22
Bangladesh, Matlab	2.38	22.93	233
Burkina Faso, Nouna	2.73	26.42	84
Burkina Faso, Ouagadougou	2.71	23.53	12
Ethiopia, Kilite Awlaelo	3.66	37.50	12
Ghana, Dodowa	3.02	26.55	77
Ghana, Navrongo	1.94	13.49	162
India, Ballabgarh	2.09	17.50	35
India, Vadu	2.59	20.00	14
Indonesia, Purworejo	2.58	25.97	20
Kenya, Kilifi	2.33	23.26	77
Kenya, Kisumu	2.35	20.10	281
Kenya, Nairobi	1.51	13.53	36
Malawi, Karonga	2.41	25.98	33
Senegal, Bandafassi	3.76	23.64	39
South Africa, Africa Centre	1.28	22.99	117
South Africa, Agincourt	2.36	21.73	216
Vietnam, Filabavi	1.87	16.88	13

From 72 330 VA of adults in 22 sites of the INDEPTH HDSS Network, 1992-2012 data

Table 9.5: Co-occurring causes by group selected through the Euclidean index of similarity (exhaustive)

Group A	Group B	Frequency	Percentages
Non-communicable diseases	**« Diseases of poverty »**	**976**	**61.3**
Diabetes and cardiovascular diseases	Infectious and parasitic diseases	648	40.7
Infectious and parasitic diseases	Other non-communicable diseases	189	11.9
Cancers	Infectious and parasitic diseases	60	3.8
Anemia and malnutrition	Diabetes and cardiovascular diseases	31	1.9
Chronic respiratory diseases	Infectious and parasitic diseases	25	1.6
Anemia and malnutrition	Chronic respiratory diseases	9	0.6
Anemia and malnutrition	Cancers	6	0.4
Anemia and malnutrition	Other non-communicable diseases	3	0.2
Diabetes and cardiovascular diseases	Maternal CoD	3	0.2
Other non-communicable diseases	Maternal CoD	2	0.1
Non-communicable diseases	**Non-communicable diseases**	**486**	**30.5**
Cancers	Diabetes and cardiovascular diseases	111	7.0
Diabetes and cardiovascular diseases	Other non-communicable diseases	132	8.3
Cancers	Chronic respiratory diseases	3	0.2
Diabetes and cardiovascular diseases	Diabetes and cardiovascular diseases	165	10.4
Cancers	Other non-communicable diseases	28	1.8
Chronic respiratory diseases	Other non-communicable diseases	7	0.4
Other non-communicable diseases	Other non-communicable diseases	1	0.1
Infectious and parasitic diseases	Maternal CoD	13	0.8
« Diseases of poverty »	« Diseases of poverty »	97	6.1
Infectious and parasitic diseases	**Infectious and parasitic diseases**	**69**	**4.3**
Anemia and malnutrition	Infectious and parasitic diseases	13	0.8
Anemia and malnutrition	Anemia and malnutrition	2	0.1
Non-communicable diseases	**Injuries and violent deaths**	**29**	**1.8**
Diabetes and cardiovascular diseases	External CoD	21	1.3
Cancers	External CoD	4	0.3
Other non-communicable diseases	External CoD	4	0.3
« Diseases of poverty »	**Injuries and violent deaths**	**3**	**0.2**
Infectious and parasitic diseases	External CoD	3	0.2

Associations irrespective of order attributed by InterVA-4

1,591 VAs of adults with co-occurring causes of death (a Euclidean index >= 0.65), INDEPTH Network, 1992-2013

Table 9.6: Association of causes (competing and co-occurring) from deaths attributed more than one cause by InterVA-4

Group A	Group B	Frequency	Percentages
Non-communicable diseases	**Non-communicable diseases**	**3004**	**38.8**
Diabetes and cardiovascular diseases	Diabetes and cardiovascular diseases	825	10.7
Cancers	Cancers	703	9.1
Cancers	Other non-communicable diseases	359	4.6
Diabetes and cardiovascular diseases	Other non-communicable diseases	358	4.6
Chronic respiratory diseases	Diabetes and cardiovascular diseases	248	3.2
Cancers	Diabetes and cardiovascular diseases	205	2.7
Other non-communicable diseases	Other non-communicable diseases	168	2.2
Cancers	Chronic respiratory diseases	81	1.0
Chronic respiratory diseases	Chronic respiratory diseases	50	0.6
Chronic respiratory diseases	Other non-communicable diseases	7	0.1
Anemia and malnutrition	Anemia and malnutrition	2	0.0
« Diseases of poverty »	**Non-communicable diseases**	**2403**	**31.1**
Cancers	Infectious and parasitic diseases	803	10.4
Diabetes and cardiovascular diseases	Infectious and parasitic diseases	670	8.7
Infectious and parasitic diseases	Other non-communicable diseases	499	6.5
Chronic respiratory diseases	Infectious and parasitic diseases	269	3.5
Diabetes and cardiovascular diseases	External CoD	60	0.8
Anemia and malnutrition	Cancers	59	0.8
Anemia and malnutrition	Diabetes and cardiovascular diseases	57	0.7
Anemia and malnutrition	Other non-communicable diseases	21	0.3
Maternal CoD	Other non-communicable diseases	10	0.1
Anemia and malnutrition	Chronic respiratory diseases	9	0.1
Diabetes and cardiovascular diseases	Maternal CoD	6	0.1
« Diseases of poverty »	**« Diseases of poverty »**	**1737**	**22.5**
Infectious and parasitic diseases	Infectious and parasitic diseases	1531	19.8
Maternal CoD	Maternal CoD	100	1.3
Anemia and malnutrition	Infectious and parasitic diseases	78	1.0
Infectious and parasitic diseases	Maternal CoD	26	0.3
Injuries and violent deaths	**Injuries and violent deaths**	**366**	**4.7**
« Diseases of poverty »	**Injuries and violent deaths**	**122**	**1.6**
External CoD	Infectious and parasitic diseases	122	1.6
Injuries and violent deaths	Non-communicable diseases	102	1.3
Diabetes and cardiovascular diseases	External CoD	60	0.8
« Diseases of poverty »	**Injuries and violent deaths**	**122**	**1.6**
Cancers	External CoD	4	0.1

Associations irrespective of order attributed by InterVA-4

7,734 VAs of adults, INDEPTH Network, 1992-2013

Table 9.7: Association with a difference between the Euclidean index based on matrix VLVL and HH >= 0.008

Cause A	Cause B	Euclidean index VLVL	Euclidean index HH	Difference VLVL - HH
Sepsis (non-obstetric)	Congenital malformation	0.66681	0.65612	0.01069
Tetanus	Congenital malformation	0.68174	0.67218	0.00955
Severe anaemia	Congenital malformation	0.70331	0.69423	0.00908
Sickle cell with crisis	Congenital malformation	0.59200	0.57922	0.01278
Liver cirrhosis	Congenital malformation	0.69001	0.68178	0.00823
Renal failure	Congenital malformation	0.68788	0.67942	0.00846
Other and unspecified NCD	Congenital malformation	0.62073	0.60796	0.01278
Congenital malformation	Neonatal sepsis	0.63779	0.62818	0.00961
Congenital malformation	Other and unspecified neonatal CoD	0.57933	0.56621	0.01312
Congenital malformation	Other and unspecified external CoD	0.69483	0.68611	0.00871

From the a priori matrices of InterVA-4

Table 9.8: Deceased by Malaria and HIV levels, HDSS

Levels	Frequency	Percentages
Low malaria & High HIV	**20715**	**28.6**
3 sites (Kenya : Nairobi - South Africa : Africa Centre, Agincourt)		
Low malaria & Low HIV	**17579**	**24.3**
9 sites (Bangladesh : AMK, Bandarban, Chakaria & Matlab - Ethiopia: Kilite Awlaelo - India : Ballabgarh, Vadu - Indonesia : Purworejo - Vietnam : Filabavi)		
High malaria & Low HIV	**17409**	**24.1**
7 sites (Côte d'Ivoire : Taabo - The Gambia : Farafenni - Burkina Faso : Nouna, Ouagadougou - Ghana : Dodowa, Navrongo - Senegal : Bandafassi)		
High malaria & High HIV	**16616**	**23.0**
3 sites (Kenya, Kilifi; Kenya, Kisumu; Malawi, Karonga)		

From 72 330 VA of adults in 22 sites of the INDEPTH HDSS Network, 1992-2012 data