A Metadata Driven Module for Managing and Interpreting HDSS Verbal Autopsy Datasets Using InterVA-4 Model

KOMBASSERE KOULIGA

STUDENT NUMBER: 1242647

A research report submitted to the Faculty of Health Science in partial fulfillment of the requirements for the degree of *Master of Science (MSc)* in Epidemiology - Research Data Management

November, 2017

Kombassere Kouliga . *2017.*
A Metadata Driven Module for Managing and Interpreting HDSS Verbal Autopsy Datasets Using InterVA-4 Model.

SUPERVISORS:
Dr. Gideon Nimako, University of The Witwatersrand
Mrs. Irma Mare, University of The Witwatersrand

SUPPORTED BY:
INDEPTH Network
University of the Witwatersrand
Ouagadougou HDSS
Dodowa HDSS

DECLARATION

I hereby declare that this research work is project work carried out by me under the guidance of Dr. Gideon Nimako and Mrs Irma Mare. I have taken care in all respect to honour the intellectual property right and have acknowledged the contribution of others for using them in this work and further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Parktown, Johannesburg, November, 2017

Kombassere Kouliga

DEDICATION

I dedicate my research work to my family and many friends. A special feeling of gratitude to my loving wife, Kombassere S. Bertille/Dombwa whose words of encouragement and her patience and support during my study and research periods. My daughter Shirley Calixte and my son W. U. Emilyo, have never left my side and are very special to me.
I also dedicate this dissertation to my many friends who have supported me throughout the process. I will always appreciate all they have done, especially for helping me develop my English skills and for the many hours of proofreading of my different projects.

ABSTRACT

Standardised by the World Health Organisation (WHO), Verbal Autopsy (VA) is a research survey-based tool widely used to interview relatives, caregivers, friends or witnesses of the deceased to collect data related to the dead in order to determine causes of death in areas where there is no medical record or formal medical attention given and deaths are not recorded routinely. In the past, the findings of the probable cause of death is usually done using the Physician Certified VA (PCVA) method where a team of physicians were used to interpret VA data in order to assign causes of death. However, this method is very time consuming and expensive in term of resources consumption. This has necessitated the need for practitioners to seek other alternate methods of determining Causes of Death (CoD). Among the several methods of collecting and analysing VA datasets, the Tariff and InterVA-4 are most widely used because they are recognised by WHO and International Network for the Demographic Evaluation of Population and Their Health (INDEPTH) Network for their and Demographic Surveillance System (HDSS) sites. The InterVA-4 is a standardised WHO verbal autopsy software used to interpret death related datasets. In this work, we have addressed some data management challenges associated with VA and InterVA-4. Among such challenges is the iterative and continual change of the WHO VA questionnaire (2007, 2012, and 2014). These set of changes come in two folds; the first changes made to the original verbal autopsy instrument by WHO to get a new version. The second usually results from each INDEPTH site adapting instrument for their HDSS area realities. Although these datasets contain spatial information such as global positioning system coordinates, the visualisation on maps of the distribution of causes of death from the verbal autopsy datasets is still lacking in the literature. In this project, we seek to fill these gaps by developing a data model(an abstract model that organizes elements of data and standardizes how they relate to one another and to properties of the real world entities) and platform for VA based on model-driven(used mostly in software design) and meta-data architectures(data about data structure and organisation). We also implemented a geographic information system (GIS) layer that allows display on maps the causes of death from verbal autopsy datasets in the demographic surveillance area (DSA). The tool will enable research scientist to better understand the patterns of causes of death in HDSS sites and aid in accurate analysis of VA datasets.

ACKNOWLEDGMENTS

I am eternally grateful to God, my creator for giving me the intellect, motivation, passion and ability to strive for better things. While a completed dissertation or research report bears the single name of the student, the process that leads to its completion is always accomplished through the support and help of many individuals and organisations. I would like to extend my sincere thanks to all of them. I am highly indebted to my supervisors, Dr. Gideon Nimako and Mrs. Irma Mare for their guidance and constant supervision as well as providing necessary information regarding the project and also for their support in completing the project.

I would like to express my gratitude towards INDEPTH NETWORK for the scholarship to enable me to undergo this Master programme at the University of the Witwatersrand of Johannesburg in South Africa. I would like to express my special gratitude and thanks to the Institut Supérieur des Sciences de la Population of Université Ouaga I Pr Joseph KI-ZERBO in Burkina Faso for allowing me to undertake my internship in this prestigious regional and international research centre in Burkina Faso. My thanks and appreciations also go to my wife Mme Kombassere/Dombwa S. Bertille for her understanding and support, and my colleagues in developing the project and the people who willingly helped me out with their efforts.

CONTENTS

ACRONYMS

ASP.NET Active Server Page

ANSI American National Standards Institute

API Application Program Interface

ANN Artificial Neural Network

CAPI Computer-Assisted Personal Interviewing

CCVA Computerised Coding of Verbal Autopsy

CSMF Causes-Specific Mortality Fraction

CRF Case Report Form

CSPro Census and Survey Processing System

CSV Comma Separated Value

CoD Causes of Death

DBMS DataBase Management System

DDI Data Documentation Initiative

DSA Demographical Surveillance Area

DIF Directory Interchange Format

GIS Geographical Information System

GPS Global Positioning System

GCP Good Clinical Practise

GUI Graphical User Interface

HDSS Health and Demographic Surveillance System

HDSA Health and Demographic Surveillance Area

HREC Human Research Ethical Committee

HTML HyperTexte Markup Language

FGDC Federal Geographic Data Committee

ICT Information and Communication Technologies

ICD International Classification of Diseases

INDEPTH International Network for the Demographic Evaluation of Populations and their Health

IHME Institute for Heath Metrics and Evaluation

InterVA Interpretation of Verbal Autopsy

ISO International Organisation for Standardisation

JSON JavaScript Object Notation

KL King-Lu

MSSQL Microsoft SQL Server

MIVA Mobile InterVA

NISO National Information Standards Organisation

ODK Open Data Kit

OpenHDS Open Health Demographic System

ORM Object-Relational Mapping

PHMRC Population Health Metrics Research Consortium

RDBMS Relational Database Management System

RDF Resource Description Framework

RDM Reference Data Model

REDCap Research Electronic Data Capture

RST Relational Schema Tier

RSP Relational Schema Protocol

RSM Relational Schema Model

SGML Standard Generalised Markup Language

SQL Structured Query Language

SSP Simplified Symptom Pattern

UI User Interface

VA Verbal Autopsy

VAS Verbal Autopsy System

XML eXtensible Mark-up Language

XHTML Extensible HyperTexte Markup Language

WHO World Health Organisation

W3C World Wide Web Consortium

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

Less than one-third (18 million) of the 56 million annual global deaths are certified through civil registration and up to 80 percent of deaths that occur outside of health facilities are not recorded or counted [18]. Such as deaths occurs predominantly in developing countries. This shows that very few developing countries have functioning cause of death information systems that they can draw on to guide policies for health programmes. The lack of reliable data on the levels and causes of mortality in disadvantaged regions of the world still hampers efforts to use reliable information, inference and indicators to support health policy, planning, monitoring and evaluation. In such scenarios, most deaths occur at home or outside of health facilities. Rather than waiting on governmental funding for well-functioning civil registration systems, most research institutions in developing countries have resulted to Verbal Autopsy (VA) as an alternate solution.

In order to assist policy makers and world organisations in reducing mortality, researchers and experts from various research centres such as Health and Demographic Surveillance System (HDSS) sites, national sample surveillance systems centres etc., are combining their efforts and experiences to harmonise various causes of death by data collection instruments and procedures available at their centres. As a result, the World Health Organisation (WHO) through consultation meetings with these scientists and experts have developed three types of standardised and harmonised questionnaires to collect cause-specific mortality data [15] specially in developing countries. These three verbal autopsy questionnaires are used to address the needs of various users including researchers, policy makers, donors and actors in monitoring and evaluation activities. These questionnaires are the result of a minuscule and rigorous work on all instruments of verbal autopsy.

The purpose of the three types of questionnaires is to allow the integration of differences in causes of death by constituting three age groups (under 4 weeks, 4 weeks to 14 years, and 15 years and above). Thus, the first questionnaire aims to distinguish between stillbirths, early neonatal deaths and late neonatal deaths and to determine the causes of these perinatal events and deaths. The purpose of the second verbal autopsy questionnaire is to determine the main causes of post-neonatal mortality in children (starting from the fourth week) and the causes of death that may have been recorded between this period and the age of 14 years old. The final verbal autopsy questionnaire identifies all major causes of death in adolescents and adults (from the age of 15 years), including deaths related to pregnancy and childbirth. Information on the symptoms and history of the disease in each age category are collected with the corresponding standard questionnaire by interviewing families, friends and intimate relatives of the deceased. This systematic approach of collection and determining the cause of death is called Verbal autopsy. Verbal Autopsy is widely used in countries where patients are rarely examined by a doctor and where medical records are inadequate or non-existent [15]. From 2007 to 2014, there have been various versions of the WHO verbal autopsy instruments and the causes of death are listed and codified based on International Classification of Diseases (ICD)-10 [29]. Although these various versions exist, questionnaire could be modified and adapted to the context and to the language by research and governmental institutions [15].

Following the collection and capture operations, the data is interpreted by at least three doctors based on their experience, training and the International Classification of Diseases. This mode of interpretation is referred to as **PCVA!** (**PCVA!**). However, given the challenges related to the high cost, slowness and non-repeatability of this mode of interpretation, other computer aided methods such as Interpretation of Verbal Autopsy (InterVA)-4, King-Lu (KL), Direct Causes-Specific Mortality Fraction (CSMF), estimate Tariff (Tariff), Random Forest (RF) and the Simplified Symptom Pattern (SSP) have been developed by some research institutions to facilitate causes of death interpretation [7]. Most HDSS sites currently use the InterVA-4 for such automated interpretation. Although InterVA-4 gives

satisfying results, it has some challenges in its use. The main challenge is the collection of the VA dataset with varying versions of the WHO VA instrument versions (2007, 2012 , 2014) [31]. Considering the size of the Verbal Autopsy questionnaire, some of HDSS sites take a lot of time to readapt or update their Verbal Autopsy tools before starting the data collection process. Due to these difficulties, some HDSS have decided not to update their VA data collection instrument. Another problem associated with this phenomenon is that, it is difficult to analyse VA dataset sampled over multiple years due to the inconsistencies in the data formats and structures. In addition to these challenges, VA datasets in most research institutions are usually managed in an ad hoc fashion by using spreadsheets. Thus, from the VA data collection to its interpretation, there is no adequate data management platform for the International Network for the Demographic Evaluation of Populations and their Health (INDEPTH) Network sites [26]. INDEPTH Network advise the HDSS Sites to use conventional relational database management models to manage and to maintain the quality of the datasets over time. The quality of datasets containing an important volume of longitudinal data from dynamic cohort follow-ups such as HDSS lies in the ability of the DataBase Management System (DBMS) to ensure that:

- the integrity and consistency of the stored data and new input data especially dates of events and episodes whose quality defines the HDSS credibility;
- the creation of good quality control procedure of data control and validation;
- control of the execution of update processes ensuring proper authorization, controlling concurrent update and synchronizing update of multiple lines;
- the data manipulation and maintenance in minimizing errors on the one hand and on the other hand to guarantee the security of some of sensitive data;
- the uniqueness of the identifiers of the individuals enlisted in the database and in the surveillance space. Each enlisted individual should not be re-enlisted with another identifier in the HDSS.

1.1 MOTIVATION

The main aim of VA is to contribute to the body of knowledge of CoD and allows policy makers and decision makers to develop appropriate strategies for healthcare programmes and governance in countries where they lack vital statistics and medical certification [3]. In the absence of appropriate tools for VA data collection and management in Health and Demographic Surveillance Area (HDSA), this aim may not be achieved [19]. This is necessary as most developing countries have HDSS sites that conduct surveillance and collect longitudinal data on vital health indicators. The design of effective public health policies and the measurement of their impact cannot be achieved without the registration of deaths and documentation of causes of death. Through the WHO standard verbal autopsy instruments (see Appendix 9.1), resource constraint countries are trying to catch up the gap of vital statistics with the HDSS sites. It is imperative that these research centres utilise an adequate data management platform to produce vital data which may be linked to similar data from other HDSSs or hospital recording systems in the same country or the region to produce more representative insight.

1.2 PROBLEM STATEMENT

The main problem being addressed in this research report lies primarily in laborious effort required by physicians in the **PCVA!** mode of interpretation. Although there has been effort to automate this process via tools such as InterVA-4, there is no data management platform for conventional VA datasets which are predominantly Comma Separated Value (CSV) files. The current VA data management processes in HDSS sites does not allow for data preservation, retention assurance and does not adhere to the suggestion made by the INDEPTH Network as well as Good Clinical Practise (GCP) [26] which is an international ethical,scientific and practical standard for designing, conducting, recording, and

reporting clinical research that involve the participation of human subjects. The second problem we address in this work is the visualisation of VA datasets. Although the data collection phase of the VA cycle collects Global Positioning System (GPS) coordinates of the subject, there is currently no tool to visualise and display the distribution of causes of death for a specific HDSS area in the context where some communicable diseases such as the Cholera exist. The VA data collection process itself is predominantly paper based. With electronic data capture penetrating in most HDSS sites, it is becoming important to digitise the data collection process of Verbal Autopsy.

1.3 MAIN CONTRIBUTIONS

The major contributions of this work is in the development and implementation of VA data platform that resolves the challenge of continual version changes and adaptation of WHO VA instruments. The main results being reported here include:

1. The implementation of a metadata driven management platform that incorporates InterVA-4 mode of interpreting VA datasets. This tool provides the data management utilities including the automatic interpretation of InterVA-4 datasets.

2. The implementation of an API that provides VA dataset feeds to Geographical Information System (GIS) applications for geospatial displays and visualisation

3. The implementation of a mapping layer that translate the metadata corresponding to the VA Case Report Form (CRF) oCRFnto Open Data Kit (ODK) eXtensible Mark-up Language (XML) data dictionary. Based on ODK's ability in data interchanging and experiences in managing metadata in the .NET environment, XML has been retained in this project to the detriment of JavaScript Object Notation (JSON). This feature enables offline VA data capture using mobiles and handheld devices in the HDSA without connectivity.

1.4 OUTLINE OF RESEARCH REPORT

The remainder of this report is organised as follows. In Chapter 2, we present the background and the related work. In the Chapter 3 we describe the metadata driven data systems module. Chapter 4 gives the metadata driven platform for verbal autopsy implemented in this project. Chapter 5 describes the experimental setup and the evaluation of the module. In Chapter 6 we give some considerations for the operational use of the platform and finally, we give the conclusion and future direction in Chapter 7.

2

BACKGROUND AND RELATED WORK

Most deaths without registration or certification occur in developing countries [23] and is usually attributed to the inaccessibility of health facilities by the population and certain cultural considerations [19]. This has resulted in the emergence of verbal autopsy, which attempts to determine causes of death for previously undocumented deceased, allowing scientists to analyse disease patterns and direct public health policy decisions. This method determines the causes of death and cause-specific mortality proportions in populations without a complete vital registration system. Data obtained from verbal autopsy comes in the form of a set of binary indicators that indicates whether an event happens or not. The process consists of a trained interviewer using the full or an adapted WHO standardised questionnaire to collect information about the signs and symptoms of events of the deceased person from his next of kin or other caregivers [10]. Then the conditional probability for each cause of death could be analysed or estimated.

Since 2007, VA dataset collected using WHO Verbal Autopsy instrument has being analysed by health professionals who assign manually a probable cause of death. This analysis method (known as the **PCVA!**) utilises the International Classification of Diseases deceases list, to assign the most probable cause of death to each case [15]. However the lack of reproducibility, the slowness and the expensive cost associated with this method has led to the computerisation of VA analysis process [5]. Thus, the WHO verbal autopsy instrument has undergone two revisions (2012, 2014) since 2007 to facilitate the use of publicly available analytical software to assign causes of death and to take into account the needs and recommendations of some professionals. These revisions also facilitate the adaptation of the questions to local context and integration. The computerisation of the verbal

autopsy data collection, interpretation and management process appear to be a challenge after the validation of the verbal autopsy questionnaire as an international standardised instrument [31]. Among Computerised Coding of Verbal Autopsy (CCVA) methods for interpreting VA data, there are on one hand algorithmic methods which follow a set of predefined diagnostic criteria giving binary outcomes (yes or no) for a single cause of death and on other hand probabilistic groups which can give the probability of multiple causes of death from a death. Among these computational interpretation methods are InterVA-4 method, King-Lu (KL) direct algorithm, Tariff method, Random Forest (RF), Artificial Neural Network (ANN) and the SSP etc.

Based on the ICD list as reference, these computational interpretation methods have been developed [7, 9] with the goal of facilitating causes of death interpretation. Among these tools, WHO recognises Tariff and InterVA-4 as analytical tools to the interpretation of VA data. Tariff [13], developed by the Population Health Metrics Research Consortium, is an additive algorithm that uses *Tariff* scores reflecting the importance and uniqueness of each symptom to each cause of death. InterVA-4 [5] is the most popular algorithm for assigning Causes of Death (CoDs) from VA data. It was developed by Umeå Centre for Global Health Research and it applies conditional probabilities of Bayesian probabilistic methods to a matrix of causes of death values and experts' opinion. The WHO standard verbal autopsy 2007 version has been adapted and standardised by the INDEPTH Network called *INDEPTH Standardised Verbal Autopsy Questionnaire* and others research centres such as Institute for Institute for Heath Metrics and Evaluation (IHME) in United States have shortened the VA questionnaire and adapted it to the local context [14, 21] through their Population Health Metrics Research Consortium (PHMRC), composed by several institutions in the world. Causes of death data in most of HDSS centres are traditionally recorded on an INDEPTH Standardised Verbal Autopsy paper-based questionnaire. Most sites have been updating their questionnaires in view of the demand of the 2014 WHO VA standard. Based on their customised and adapted VA questionnaire, the PHMRC have developed a verbal autopsy data collection and interpretation tool called Tariff. This tool

is not used by most of INDEPTH Network sites because their adapted VA questionnaire is different from the one recommended by INDEPTH.

Agincourt HDSS in South Africa has started computerising the data collection of causes of death using their adapted Verbal Autopsy questionnaire with mobile phones [27] to allow for the Computer-Assisted Personal Interviewing (CAPI) method. This computerised tool, (called Mobile InterVA (MIVA)) [27], does not permit the dynamic creation of VA questionnaire forms which can accommodate frequent changes and the local adaptability of WHO VA questionnaire. It also does not have any visualisation utility to display the distribution of causes of death on maps. The tool does not provide automatic causes of death interpretations and does not use the data model suggested by the INDEPTH Network. According to the InterVA-4 and Tariff documentations, the input data to allow causes of death interpretation must be in CSV format. As a result, most research centres stores VA data in CSV format. This does not allow preservation of the contextual metadata and also does not support recovery. That is why in the longitudinal studies with a dynamic cohort, such as the HDSS, the Reference Data Model (RDM) [4] is recommended by INDEPTH Network to manipulate, manage and preserve the quality of the data collected over extended periods of time [26] in a defined area.

All research centres that are engaged in longitudinal data collection on population have the common demographic core structures and processes. This core structure is based on the entry and/or exit of the individuals in the HDSS defined geographical area. In the monitoring of the presence of these individuals, HDSS need to keep the data of its population whose composition varies by birth, death and migration. In addition, in order to be able to manage the operation in the field and the cleaning of the routine data, the place, the author and the date of the collected events must be kept. One of the other big challenges is the management of dates which is one of the fundamental data element of the HDSS data for credibility and quality assurance. There must be a strong consistency between dates of events and episodes [20] in HDSS data. An event is a fact that indicates a change in the state of an individual enlisted and followed in the defined surveillance area

of an HDSS. Among these events, there are deaths, births,in and out migrations, changes in marital status. While episodes are significant and identifiable segments or time intervals in an individual that begin and end with events. As an illustration, the life of an individual, can be taken as an episode that began with his birth and ended with his death. The design of a database for an HDSS is enamelled by numerous difficulties in view of all these complex requirements. Given these difficulties, a reference data model has been designed under the leadership of INDEPTH Network to assist HDSS sites in their database design and the exchange of data. This should be recommended as integral to the reference model.

INDEPTH encourages best practices for data management by using this reference data model based on the relational database model and incorporating it into a HDSS 'resource kit' available on their website [22]. The database design of most HDSS is based on this reference model. However, this is static model although it allows for a consistent data system. This makes it difficult to add new tables to the reference model. The lack of proven technical skills in the database design for the implementation of these types of operations could also have a negative impact on the quality of the data in general. If it is true that the basis for obtaining quality data lies in the quality of the datasets entered by field workers, the development of data entry screens is also a fundamental element in the data management life cycle.

This good practice of data entry form design has an impact on the HDSS database design and data store through the data reference model. The limitations of the data reference model proposed by INDEPTH for HDSS comes from taking this good practice into account. Given that all research centres that perform longitudinal surveillance on a population have a common demographic core structure, a certain number of data classes or response modalities that have many homonyms and/or synonyms could be taken into account by this reference data model.

Prior to the development of Open Health Demographic System (OpenHDS), these common data classes were not taken into account in the HDSS databases design. They were hard-coded in the development of the longitudinal data collection tools of HDSS. Verbal

autopsy questions can change according to the researchers needs (addition or suppression of response modalities) and some local specificities (local languages, local ethics settings and behaviours). These changes are made manually and are not usually recorded or kept within the HDSS database or any other database to allow track changes. As a result, the goal of creating a standardised model for data sharing within the INDEPTH Network and helping research centres to develop accurate and reusable software [4] is not well achieved. Given these insufficiencies, the data models used vary from one site of the INDEPTH Network to another with practices that do not guarantee the quality of the data. However, one thing is to take into account these best practices in the data reference model, but another is to propose a flexible model that would really allow data managers to develop a precise and reusable dynamic system for all the INDEPTH Network sites. This has inspired the use of OpenHDS [12] which has a flexible and metadata driven model based the static INDEPTH reference data model. It is a precise but dynamic platform allowing each HDSS to adapt it to these data management challenges. Although OpenHDS has an integrated form generation utility for the verbal autopsy data collection, it does not have tools for the interpretation of these dataset. It also does not have utilities that allows the geospatial visualisation of the distribution of cause of death. Hence the importance of this research project, which entails the implementation of this dynamic data management module for VA data management based on the standard VA instruments and InterVA-4 model. This platform caters for the frequent changes and the local adaptability of WHO VA questionnaires.

The related works presented in this chapter shows that there is a paucity of tools allowing data capture form generation for verbal autopsy data capture in the INDEPTH Network research centres. In a context where epidemics such as the cholera has caused many deaths, there is a need for verbal autopsy tool to provide instruments which can automatically identify causes of death, their location and the population at risk around that location. Our tool enables data capture form generation, verbal autopsy data interpretation using InterVA-4, the visualisation of the location of the diseases and distribution

of CoD. This tool is an effort that will allow researchers and policy makers to plan and design appropriate intervention programmes to curb communicable and non-communicable diseases in resource limited communities.

3

SOFTWARE DEVELOPMENT

The advent of the internet has fundamentally changed the tools in daily activities and has also driven emergence of web applications. Most of these applications have backend relational database or key-value stores. In recent years, there has been increased demand for applications that allow modification of the underlying data structures with no programming effort. Users may perform actions that result in redefining the underlying schema [14]. This requires data systems that enable schema alteration in real-time, i.e. without recompiling the application. As such there is a need for new system development techniques for data applications. In this work, we employ a metadata-driven approach to relational database design as a solution to this problem. These new techniques and approaches to data system designs reduce the time and effort the developers spend on programming during system upgrades and updates [28].

Metadata is data about data or information about information. It summarises basic information about a dataset, which enables working with particular instances of data easier. Metadata is defined by the National Information Standards Organisation (NISO) as a "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource" [11]. Metadata describes the properties and capabilities of connections between different types of information in a particular application domain such as applications, databases, or a simple processing engine.

The various types of metadata include descriptive metadata, structural metadata and administrative metadata. Metadata is used for describing formal schemes of resources that allows the searching of resource items through keyword criteria and regrouping of the location of different resources. Most metadata schemes or syntaxes are expressed in

a number of different markup or programming languages such as HyperTexte Markup Language (HTML), XML, etc., each of which requires a different syntax to structure the metadata content. Nowadays, most metadata syntax is defined in XML. In the major DBMS, metadata is usually described as names, sizes, and other properties of database objects like tables, columns, primary key, foreign key references, data types, etc.

With the technology evolution, International standards such as American National Standards Institute (ANSI) and International Organisation for Standardisation (ISO) have been created and applied to metadata to reach consensus on standardising metadata. Thus, many current schemes use Standard Generalised Markup Language (SGML) or XML to create the metadata content. Based on SGML, many different metadata standards schemes have been developed across disciplines, including the Data Documentation Initiative (DDI), the Dublin Core , the Directory Interchange Format (DIF),EBUCore metadata,the Resource Description Framework (RDF) and the Federal Geographic Data Committee (FGDC) . Witten in XML by the membership-based Alliance, the DDI is a widely used, international standard for describing data from the social, behavioural, and economic sciences. The Dublin Core, sponsored by the Dublin Core Metadata Initiative and published as ISO Standard, is a basic, domain-agnostic standard which can be easily understood and implemented, and as such is one of the best known and most widely used metadata standards. Sponsored by the Global Change Master Directory and defined as a World Wide Web Consortium (W_3C) XML Schema, the DIF is an early metadata initiative from the Earth sciences community, intended for the description of scientific data sets. Adapted to audio-visual content, EBUCore is a set of descriptive and technical metadata based on the Dublin Core. Although not a standard, Microformat is a web-based approach to semantic markup which seeks to re-use existing HTML/Extensible HyperTexte Markup Language (XHTML) tags to convey metadata. While the FGDC is a metadata standard developed to determine the robustness, the method of accessing, and successful transfer of geospatial resources. Developed under the auspices of the World Wide Web Consortium the RDF is an infrastructure that enables

the encoding, exchange and reuse of structured metadata based on the application of XML.

Used within the computer science community, metadata served in recent health research software development such as Research Electronic Data Capture (REDCap), Census and Survey Processing System (CSPro), ODK etc. Metadata driven applications are vital in scenarios that requires dynamic content generation such as VA questionnaire CRF. For instance, with such design, data managers can control all the rich behaviour of form fields, including formatting, validation, visibility, User Interface (UI) type (drop-down vs text field vs radio buttons, etc.). It also allows single application models to generate many different variations based on data structures that are defined at runtime in a database containing metadata which describes the fields of the application. CRFs and other UIs are not programmed but are automatically generated, based on metadata stored in a database, including variable name, variable type, question prompt etc. Changes or additions to generated forms are managed in metadata through the database. Thus, instead of repeatedly developing the same type of form or interface for web application with similar needs, there is only one program that works on different sets of metadata for different users. Since metadata is also data, it is easier to manage form creation elements stored as data in the application relational database than to manage source code for web application programs.

We utilised such a metadata driven development model to enable the re-adaptability of the autopsy verbal questionnaires in HDSS research centres. The use of the metadata data modelling facilitates the changes of the CRFs, without each adaptation or change of the different verbal autopsy questionnaire requiring modification of source codes. This made it possible to recreate or make the necessary modifications in the creation of the VA forms without any modification in the application layer. Our solution achieved more than tools such as REDCap, ODK, CSPro etc. could have, as it managed the dynamic modes of VA CRFs as well as the interpretation and visualisation of verbal Autopsy datasets. In addition, our tool has utilities that allows the use of mobile technologies. Primarily this

utility translates the generated HTML form to XForm format. This XForm format is then used by ODK Collect for offline or online Verbal Autopsy data collection using mobile and handheld devices.

3.1 METADATA DEVELOPMENT MODEL

The evolution of Internet-related technologies and the exponential use of all its tools have forced most web application infrastructure providers to frequently update and maintain their products in order to remain competitive in an aggressively evolving technological environment. These frequent updates and maintenance actions lead to a significant increase in the cost of web applications development. In addition, the development of software multitenancy is much more complex than single client applications, requiring highly skilled software developers and increased development lead times. As such, application developers are recently increasingly using the metadata driven approach as a solution to automatically and dynamically generate Graphical User Interface (GUI) and relational database schema that meet user demands. This reduces work effort, cost and time of application development and increases the independence of users and developers. From the literature, the approaches developed to generate on the fly graphical user interfaces and database schema can be summarised in two main approaches: The first approach is the model where graphical user interfaces are generated from metadata stored within a database; the second one concerns the approach where metadata information is captured using the user interface to automatically create and/or update database table information on the fly. However, before giving more details on these two approaches, it is important to describe metadata in detail.

3.1.1 *Auto-Configuration and Modification of Relational Model using Metadata*

From relational databases to object databases, Structured Query Language (SQL) has remained the common language for querying database management systems. The use of this language, which is mainly used for the definition and manipulation of data into databases, requires the knowledge of the relational schema of the database. The database structure is described by metadata tables which include the four essential elements required to represent a data model: table names, field names, field data types, and linkages among the tables. The link definition requires two fields to specify both the linked table and the linked field within the table. The metadata table also stores descriptive information that helps the user to understand the meaning and appropriate use of each data field in the data dictionary(See Appendix).

Since relational schemas are usually customised for application logic, the data model evolution involves either redefining the SQL queries through hard-coding during the development/maintenance of the application, or regenerating it using a rational or an Object-Relational Mapping (ORM) tool. All the modules or codes of the concerned applications which are in interaction must be implemented and compiled again. The user is usually subject to the whims and requirements of the developer or supplier of the database. Such maintenance or upgrades can be time consuming because in certain scenarios, it is necessary to stop the database in production for the adjustments of the schema. In other cases, the adjustment operation of the current relational schema might not be possible if it has already undergone many adjustments, making it unusable.

In order to avoid recompiling or rewriting source code when upgrading or updating relational schemas, an application must be able to retrieve database schema information in order to automatically generate or update SQL queries into the current application. The weakness of the metadata of the relational schema resulting from the views of the DBMSs's information system , the inextensibility of this relational schema and its incompatibility

from one DBMS to another among others, present some challenges [25] in the implementation of the solution of this project. It is almost impossible to map database tables to the internal classes and objects of the DBMS's information system relational schema at the application level to allow the automatic generation of the user interface from the existing relational schema [25]. Facing to these challenges, Msgr. Vojtěch Přehnal proposed an alternative solution through a new relational database metadata model that would facilitate automated management of relational schemes [25]. In addition to this schema, a new communication protocol has been developed to allow remote and independent procedure calls between the relational schema and the user application [25]. The implementation of these new tools and techniques allows the representation and storage of relational metadata, the exploration of the relational schema and the exchange of data and metadata. Based on the standard (existing) relational metadata model, the revised structure of the new model that contains additional metadata for relational schema localisation, data visualisation and validation allows for more efficient processing of metadata from the information schema of databases.

In the proposed model presented in [25], relational metadata is stored in regular database tables instead of being extracted directly from the database engine information schema views. In addition, access to the database cannot be done using SQL commands, but through a proposed Relational Schema Tier (RST). The exchange of relational data and metadata is done with this level through the proposed Relational Schema Protocol (RSP), and according to metadata changes this level automatically changes the relational schema.

Figure 3.1 illustrates the proposed approach where the Relational Schema Model (RSM) represents relational metadata. Instead of being extracted from the information schema of the database engine, these relational metadata are stored directly in the regular database tables which are accessible using a new software tier called RST and not through SQL commands. This allows the realisation of exchange using XMLlanguage between the rela-

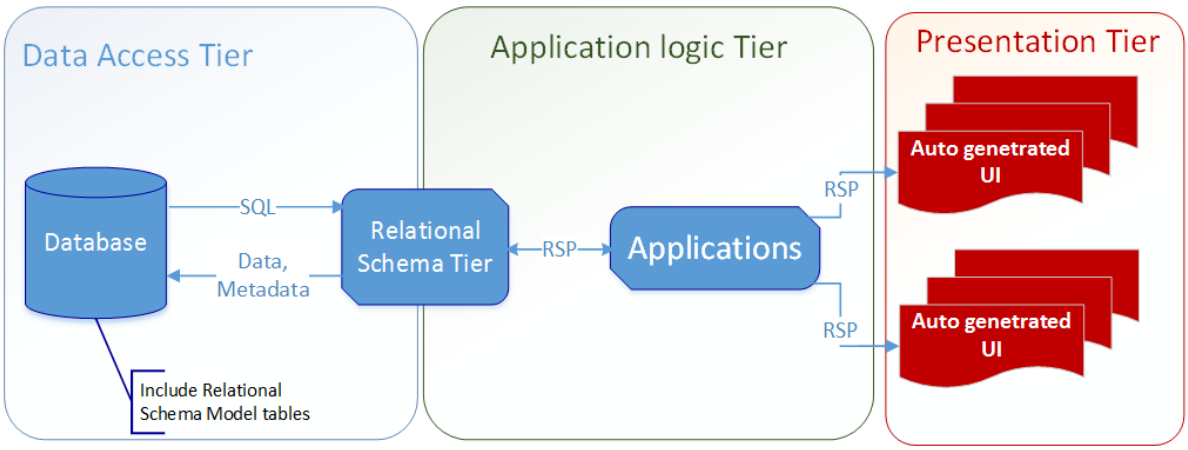

Figure 3.1: Architecture of the Proposed Approach (From [25] with some adaptations)

tional data and the metadata and favours, according to changes of metadata, an automatic modification of the relational schema.

3.1.2 *User Interfaces Generated from Database Metadata*

This approach allows the end user to automatically generate and dynamically manage user interfaces from the database without any script or coding. Given the changing needs of users in the use of Internet services, dynamic interaction between databases and the web resources has become the cornerstone of web application development. The variation of needs from one client to another requires interfaces that are extensible. The UI business logic of client-side may also need customisation based on individual end-user needs. A data entry form as user interface might also be different from user to another user. This may include fields and controls position, field's type linked to data type, etc. Client-side business logic customisation also includes customising validation rules, changing control properties, and other modifications. Developing user interfaces from scratch or based on existing applications make web application user interface development as a waste of resources that could be better spent on others domains. Hence the emergence of techniques that enable the extension or customisation of web applications according to new needs. Most applications solve this problem by storing customisable items such as

UI layout, questionnaire fields and client-side business logic as metadata in a repository. The metadata can then be interpreted by a run-time engine to automatically and dynamically generate and display the form or screen to end-users and to execute the client-side business logic when the user performs an action on the screen.

The benefits of this approach in threefold: 1) There is an absence of a recompilation of the application code and the redeployment of components on the presentation layer because the customisation is done in a central repository; 2) It only requires a very light client installation by deploying only the runtime on the client machine; 3) In addition to reducing the time and cost, and allowing an application sharing to multiple users, this approach permits users who have basic computer skills and no skills in application coding to be able to integrate their new needs without typing any code.

The implementation of metadata driven user interfaces requires three design elements. The first is to design the metadata relational schemas and decide on the repository storage mechanism. Repository storage could be a Relational Database Management System (RDBMS) such as Microsoft SQL Server (MSSQL), as is the case in this project, or any other data store format such as XML files.

In Figure 2, the metadata driven process schema and overview of the design process are presented. The designer designs and develops a graphical user interface through metadata tags that are stored in the relational database and which allow the end user to dynamically create and automatically generate a user interface by using web APIs. The main data used to create and generate the user interfaces as well as the data entered with the generated interface are stored in a single database.

Figure 3 shows an arrangement of the relationships between the tables in the relational database as well as the raw data of the dynamic and automatic user interface generation project. The four examples of tables in this schema show the association of these tables. The first table stores the data for each user interface project created by the end user. Several objects can be related to a project. Each row of these objects in the object or content table can contain multiple controls such as text fields, combo boxes, radio buttons etc. This

diagram gives an overview of what might be a relational database schema for setting up a metadata driven application for automatic user interfaces generation from database metadata.

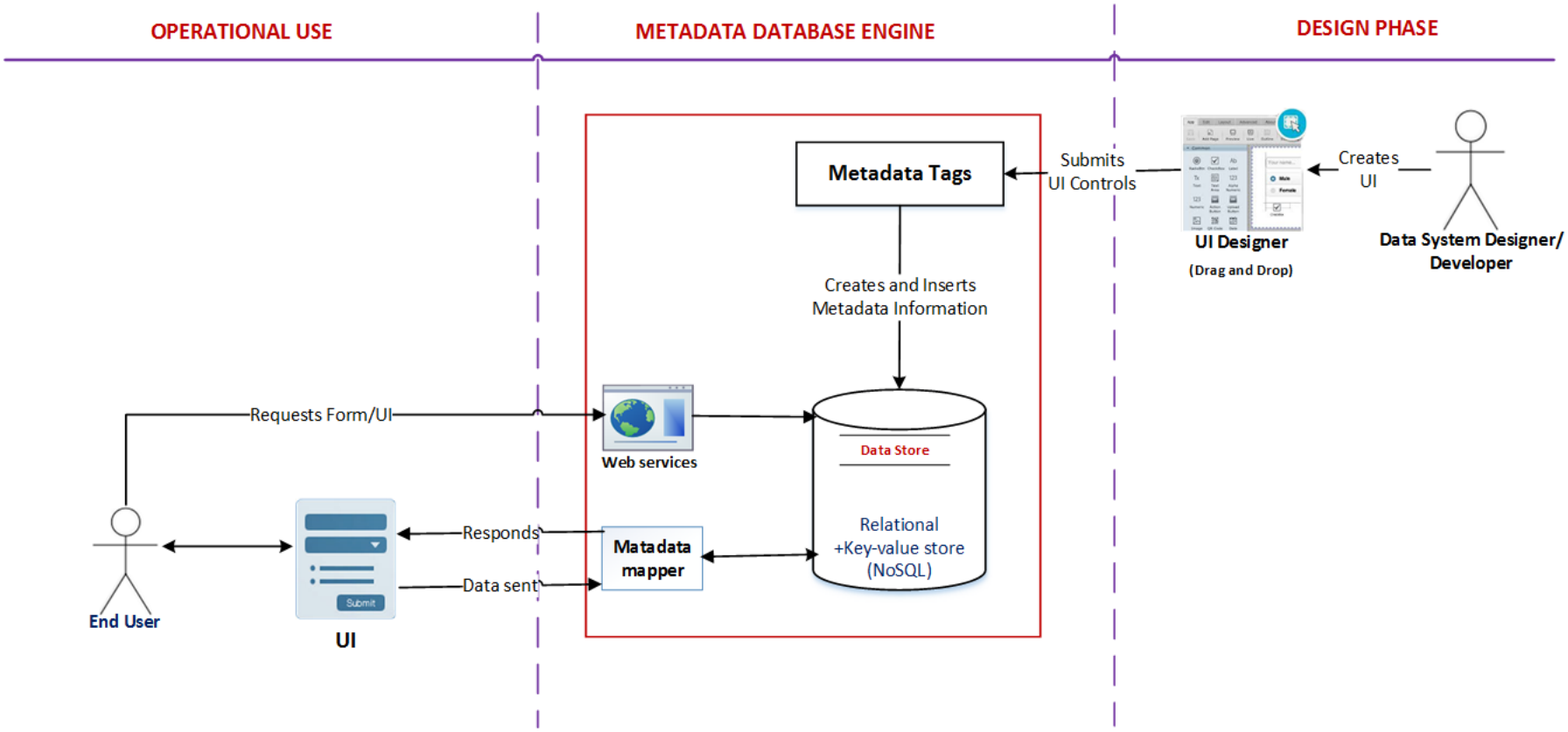

Figure 3.2: Metadata Driven Process Flow

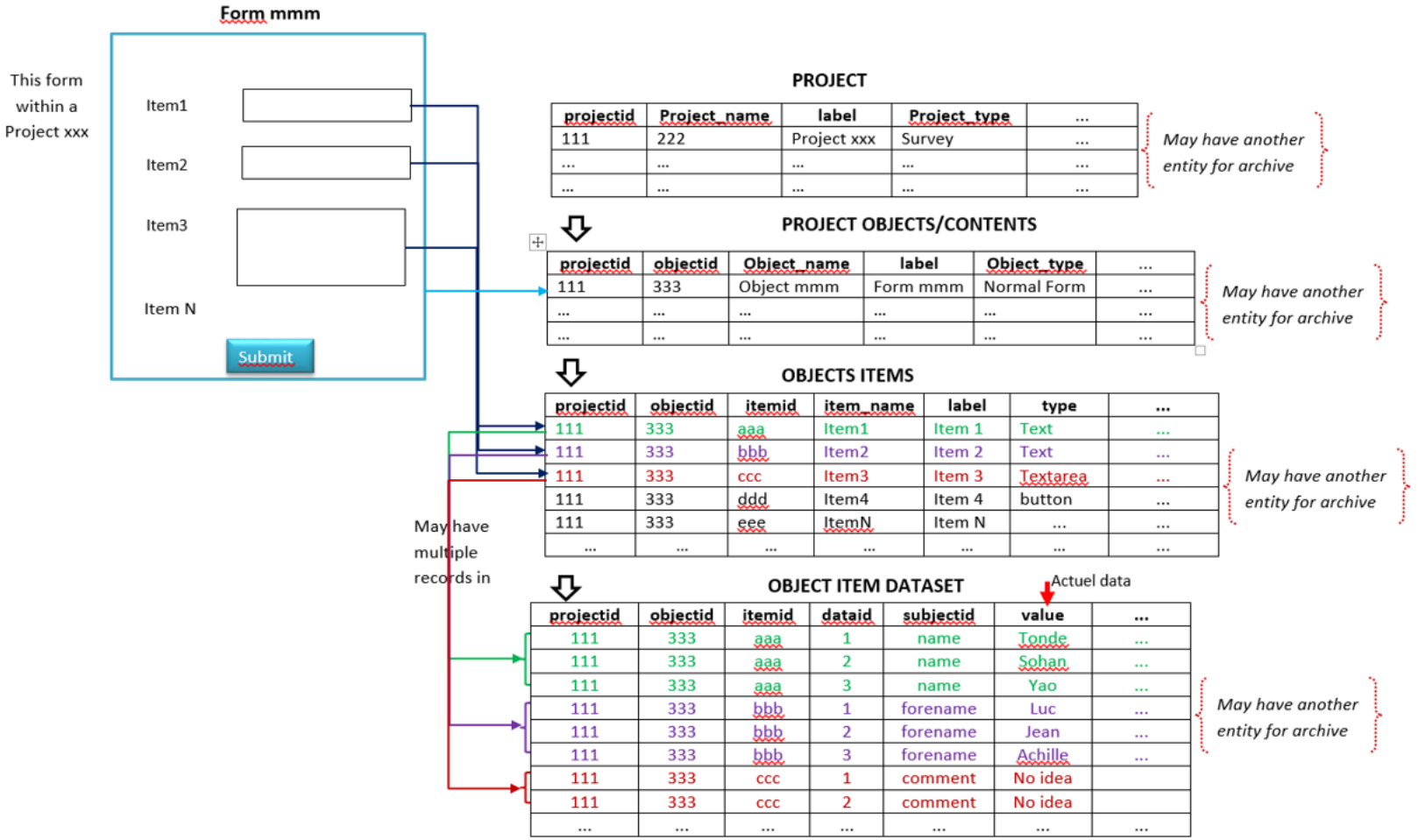

Figure 3.3: Generic Metadata Structure

3.2 OPEN DATA KIT

Over the last decades, advances in Information and Communication Technologies (ICT) have transformed the way we create, retrieve, update, and delete information. However, much of the world do not benefit of these technological advancements. As such there has been a push from development agencies to apply evidence-based development in technology. To address technological challenges faced by researchers in developing countries, ODK [1], a modular, extensible, and open-source suite of tools designed to empower users to build mobile information services was developed.

ODK enables fast and efficient online or offline data collection on mobile devices and let users own, visualise, and share data without the difficulties of setting up and maintaining servers. ODK currently consists of four utility tools namely *Collect*, *Aggregate*, *Voice*, and *Build*. ODK Collect is a mobile platform that renders complex application logic and supports the manipulation of data types that include text, location, images, audio, video, and bar-codes. ODK Aggregate provides a click-to-deploy service that supports data upload, storage and transfer in the cloud as well as on local servers. ODK Voice renders application logic using automated phone prompts that users respond to with keypad presses. Finally, ODK Build is a drag-and-drop application designer that generates the logic used by the tools.

It is anticipated that the translation of the form generated by our module to ODK will aid HDSS sites to be more efficient in VA data capture and its interpretation. Currently ODK is being used in major HDSS sites for other cohort studies.

3.3 VISUALISING CAUSE OF DEATH OF VERBAL AUTOPSY DATASETS

Public health is a growing field that is in recent time using GIS for research applications. GIS offers healthcare professionals the ability to identify health related trends and more

thoroughly target health interventions based on locality. With the use of geography and other inputs as well as requisite mathematical models, GIS can identify where diseases are most likely to spread next [16]. In field-based epidemiology, surveillance data collected usually contain geographical coordinates that can be used for spatial modelling [24]. For HDSS sites, GIS can aid in improving the public health surveillance and monitoring of diseases as well as facilitating decisions making and interventions. GIS in public health is usually associated with three main epidemiological variables, namely time, personal identifier and place. The place (usually GPS coordinates) has always been the most difficult to analyse and interpret [2]. Although the use of GIS in the public health sector faces some obstacles such as the protection of privacy and confidentiality, the capabilities of GIS to link spatial datasets to disease spread/trends offers significant and direct benefits to populations at risk. The realisation of this possibility at research institutions where longitudinal surveillance data as well as VA data are collected over time can contribute and strengthen epidemiological research and also improve the dissemination of data through visualisation to facilitate decisions making and interventions.

Africa has more than 44 HDSS sites according to the INDEPTH network [30]. Most of the studies on VA datasets compare **PCVA!** and computerised methods in verbal autopsy data interpretation, and attempts to identify risk factors for leading causes of death without taking into account the place where these causes of death can be found [5, 27] . We implemented a visualisation layer on top of our VA data management platform to address this gap.

4

PLATFORM FOR VERBAL AUTOPSY

In this chapter, we give a description of the data repository platform for verbal autopsy based on metadata driven system design and development. We outline a scope of the system as well as some ethical considerations.

4.1 PROTOTYPE DESIGN

The main component of our platform is based on the implementation of the metadata design frameworks presented in Section 3.1. This system replaced the ad-hoc tools used by most HDSS sites for verbal autopsy data capture and interpretation. All the processes, from data collection or entry, interpretation and management, are done using one platform. We first give an overview of how we employ metadata driven development methodology described in Section 3.1 to implement this platform.

4.1.1 *High Level Design*

Figure 4.1 illustrate the use-case diagram of our platform. This diagram is composed of five use cases:

1. *Create Data Entry Form allows* any user to create his own data entry form for Verbal Autopsy questionnaire through an entry utility that receives the questionnaire metadata.

2. *Display Causes of Death Repartition* permits any user to request to the system the distribution of causes of death by verbal autopsy death broad categories (Communicable, non-communicable and injuries diseases) of the HDSS area

3. *Find Probable Causes of Death* allows the user to get the probable CoD results and interpretation

4. *Entry Verbal Autopsy Death* describes the processes of the verbal autopsy data entry based on the data collected on the reasons related to the death. The entered data are interpreted using the use case 2. *Find Probable Causes of Death*.

5. *User Authentication* explains the user connection to the system during the use or running of one of the use cases. The Appendix A.1 gives more details on theses use cases.

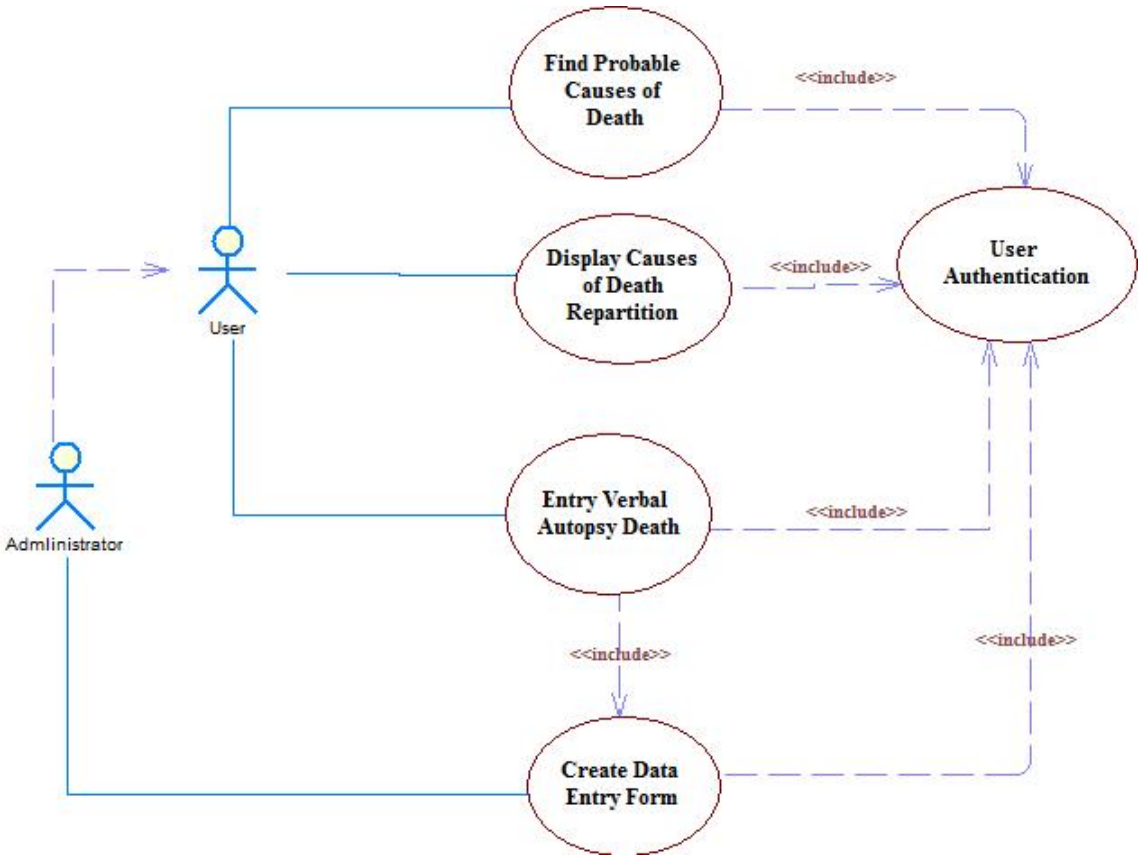

Figure 4.1: Use Case Diagram

Based on these use case actions and the generic metadata structure, a relational database model (see Appendix A.2) and an architecture has been created and presented in this report to give an overview of the system.

4.2 OVERVIEW OF THE VERBAL AUTOPSY DATA MANAGEMENT PLATFORM

4.2.1 *Architecture*

The architecture gives a description of five main components which have been implemented in the platform. The first component deals with Verbal Autopsy data collection instrument generation and the second allows Verbal Autopsy entry whiles the third (ODK data dictionary) component handles the translation of the CRFs forms to ODK formats. The fourth component manages the interpretation of the verbal autopsy entered data using InterVA-4 algorithm which are based on Bayes' theorem [5]. The final component displays the CoD distributions grouped according to the cause of death board on maps. In Figure 5, we use the green rectangles to denote these functionalities and components. The last part of this architecture concerns the deliverables of the platform to the end users. These deliverables of the module comprise of maps, basic statistics, CoD interpretation and data ready for statistical analysis in viewable and exportable formats like png or csv. Based on this architecture below, a platform has been implemented to allow verbal autopsy data management and interpretation on a unique area for health and demographic surveillance sites.

4.2.2 *VA Form Generation and Data Entry*

Verbal autopsy questionnaire has 3 components: *Adult*, *Child*, and *Perinatal*. The questions' variables are already codified by the WHO group experts. Considered metadata, most VA questionnaire questions have codes in the following format: *3A260*, *3A270* and *3A280*, etc.

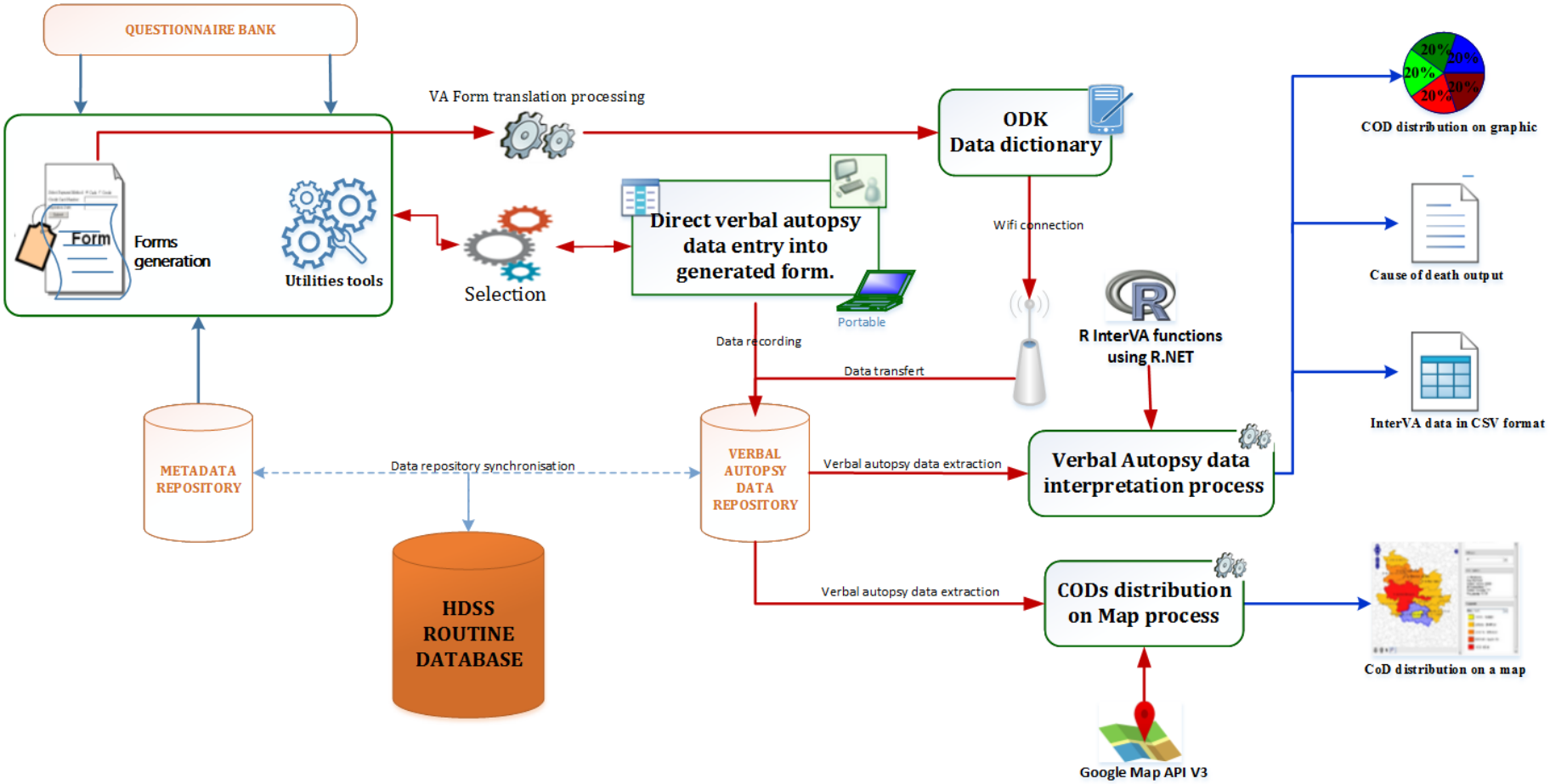

Figure 4.2: Architecture of VAS

Each code corresponds to a question on the VA questionnaire. In situations where a question item has more than one answer, its code is indexed such as 3A260A, 3A260B, etc. The set of questions from all autopsy verbal questionnaires are organised and coded in chronological order. That is, when a question is repeated in the three types of questionnaire, this question has the same code everywhere.

In the process of generating forms through this platform, the user must use the existing codes of verbal autopsy questions and upload them through a form dedicated to the entries of metadata (see Figure 4.3). A `RenderHtml()` function takes as arguments the codes in addition to other metadata such as question answer types, question labels etc., to automatically generate the form for entering verbal autopsy data. This function returns a string containing all of the form's html codes or tags. The code from the generated form is in XHTML format. The translation of this XHTML format to XML allows compatibility with XForm to be accepted by ODK Collect API.

Figure 4.3: Questions Metadata Entry Form extract.

The 246 variables of the InterVA-4 matrix are extracted from the verbal autopsy data entered through the platform. Since the verbal autopsy variables are not identical to the variables of the InterVA-4 matrix, the InterVA-4 User Guide (version 4.RC1 2012-08-14) was

used as a reference to automatically map these two variables set. The InterVA-4 matrix is also composed of questions from the three types of verbal autopsy questionnaire with much more precision on the types of responses. So, all the variables in the InterVA-4 matrix are closely related to all the questions emanating from the three types of autopsy verbal questionnaire. This has been taken into account in the codification of the verbal autopsy questions, in order to facilitate the link between the questions of verbal autopsy and those of InterVA-4. As a result, the three types of autopsy verbal questionnaire have been coded in a single, coherent order to save time needed for data extraction.

4.2.3 *Interpretation of Verbal Autopsy Data*

The VA data interpretation precedes the cleaning of these data. This cleaning processes are managed by stored procedures based on predefined logic. Once the data is cleaned, the VA data interpretation can start using the cleaned data with InterVA-4 function of R. The InterVA-4 function takes as input the VA data as matrix and other specified model parameters. The InterVA-4 function which is written in R has been converted through R.NET to allow its use with .NET code. Thus, this function of R used in this project through R.NET is:

```
1: procedure INTERVAFUNCTION
2:     HIV ← '0'
3:     Malaria ← '1'
4:     directory ← dirPath
5:     filename ← '2'
6:     output ← '3'
7:     append ← FALSE
8:     groupcode ← FALSE
9:     replicate ← FALSE
10:    if filename = path then
11:        VAOUTPUT ← InterVA(HIV, Malaria, directory, filename, output,
           append, groupcode, replicate).
12:        stringformat ← string.Format(VAOUTPUT, hivp, malariap,
           filename, output).
13:        result ← engine.Evaluate(stringformat).
14:    return result
```

The input of VA (i.e., `VAINPUT`) data is automatically received by a function that extracts the dataset directly from the underlying database. Among the parameters [17] of the InterVA-4 function, the most important are:

- `hivp (HIV prevalence)(0)`
- `malariap (malaria prevalence)(1)`

Others are optional:

- `directory, filename (2)`
- `output (3)`
- `append`
- `groupcode`

- `replicate` etc.

The `(0)`, `(1)`, `(2)` , `(3)` represents the values of the function parameters such as `hivp`, `malariap`, `filename`, `output`, in which their values are given by the end user. For this module, the function has to receive these important parameters from a dedicated form where some optional parameters are added to create an input value set.

The InterVA-4 function of R is executed through the system on the extracted datasets to obtain the CoD for each deceased. All the content of this result except the values of the variables `ID`, `MALPREV`, `HIVPREV` automatically generated by the InterVA-4 function can be exported into Excel 2007 and earlier version (see Figure 4.4). The content of the `ID`

InterVA Interpretation results for each death

ID	MALPREV	HIVPREV	PREGSTAT	PREGLIK	PRMAT	INDET	CAUSE1	LIK1	CAUSE2	LIK2	CAUSE3
155	I	I	Indet	0			Acute resp infect incl pneumonia	75			
212	I	I	nrp	99	1		Malaria	88			
222	I	I	Indet	0			Acute abdomen	65	Congenital malformation	32	
328	I	I	Indet	0			HIV/AIDS related death	98			
341	I	I	Indet	0			Malaria	47			
372	I	I	Indet	0			Diabetes mellitus	100			
678	I	I	Indet	0			Malaria	98			
907	I	I	nrp	100	0		Severe malnutrition	97			
1006	I	I	Indet	0			Acute resp infect incl pneumonia	100			
1144	I	I	Indet	0			Malaria	67			
1161	I	I	Indet	0		Indet					
1167	I	I	Indet	0			Malaria	77			
1237	I	I	Indet	0			HIV/AIDS related death	97			
1262	I	I	Indet	0			Meningitis and encephalitis	95			
1294	I	I	nrp	63	37		Malaria	84			
1327	I	I	Indet	0			Other transport accident	91			
1363	I	I	Indet				Malaria	100			

Figure 4.4: An Extract of Verbal Autopsy Data Interpretation Result

variable is an anonymised individual ID from the HDSS routine database. The content of the variables `MALPREV`, `HIVPREV` is the prevalence of Malaria and HIV given as parameters during the execution of the InterVA-4 function.

Figure 4.5 shows the cause-specific mortality fraction and the probability of the cause of death distribution in the population. These results constitute the final results of the verbal autopsy data interpretation using the InterVA-4 function of R.

Cause-Specific Mortality Fraction from Verbal Autopsy Data

Causes of Death	Cause-Specific Mortality Fraction
"01.01 Sepsis (non-obstetric)"	0.00762829926001232
"01.02 Acute resp infect incl pneumonia"	0.13573393817537
"01.03 HIV/AIDS related death"	0.127963659967164
"01.04 Diarrhoeal diseases"	0.0691582748087491
"01.05 Malaria"	0.263652900887756
"01.06 Measles"	6.95352124565253e-05
"01.07 Meningitis and encephalitis"	0.0711930689314201
"01.08 & 10.05 Tetanus"	2.57885417640089e-06
"01.09 Pulmonary tuberculosis"	0.00648467655240719
"01.10 Pertussis"	1.5945057026797e-06
"01.11 Haemorrhagic fever"	0.0158007297764794
"01.99 Other and unspecified infect dis"	0.0056367649826351
"02.01 Oral neoplasms"	4.6368969890351e-17
"02.02 Digestive neoplasms"	3.70951759122808e-14
"02.03 Respiratory neoplasms"	1.15922424725877e-16
"02.04 Breast neoplasms"	1.85475879561404e-13
"02.05 & 02.06 Reproductive neoplasms MF"	1.85475879561404e-15
"02.99 Other and unspecified neoplasms"	9.52506640125495e-05
"03.01 Severe anaemia"	6.44668833169309e-06
"03.02 Severe malnutrition"	0.0488921077957805

Graphical Representation

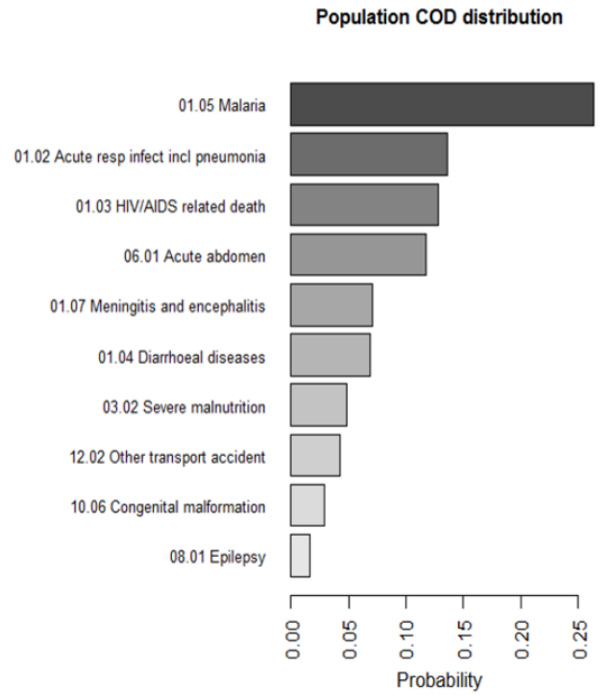

Figure 4.5: An Extract of Cause-Specific Mortality Fraction and the Probability of the CoD Distribution.

Our tool can also indicate the probability of people to die of a specific disease such as Malaria, HIV/AIDS, Server Malnutrition, etc. In Figure 4.6, such output is used to show that the probabilities for Malaria and acute respiration infection in HDSS area are higher than the rest of the causes of death. All the statistics accompanied by charts are automatically generated from the VA entered data by this metadata driven module.

4.3 PROBABILISTIC cod! (cod!) DERIVATION

The deduction of a cause of death from a predefined set of causes consistent with the International Classification of Diseases Version 10 (ICD-10) is made with the interVA-4 software and the InterVA-4 R package. The InterVA-4 R package kernel has been implemented based on the model underlying InterVA-4 as published by Byass et al., and uses

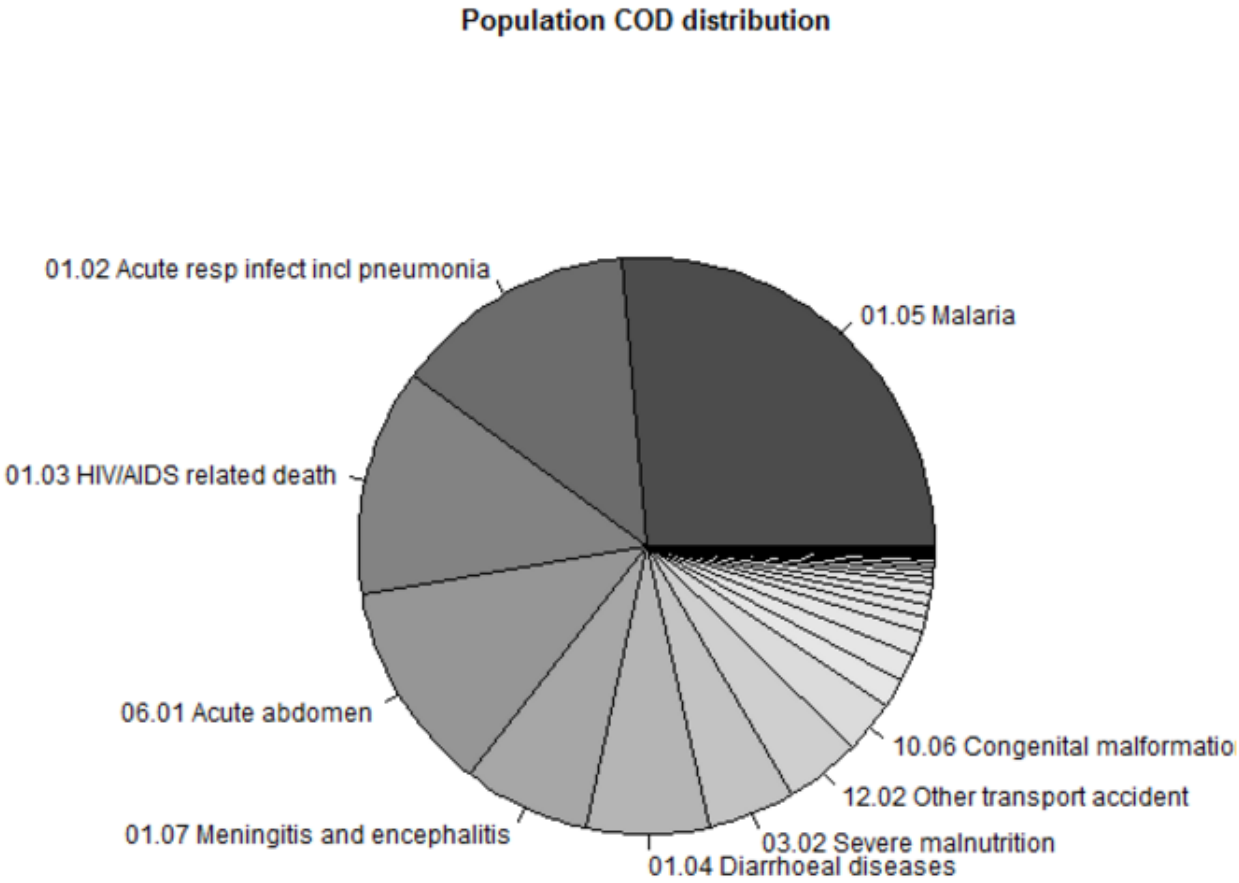

Figure 4.6: Population of CoD distribution

the Bayes theorem to determine the conditional probability of each given cause of death as a function of a set of events [17]. The events in this context are the signs, symptoms and circumstances listed in the interview questionnaire whose data are obtained from the verbal autopsy in the form of a set of binary indicators representing whether the event occurs or no. Thus, the conditional probability for each cause of death could be calculated as follows:

$$P(C_i|I) = \frac{P(I|C_i)P(C_i)}{P(I|C_i)P(C_i) + P(I|(1 - P(C_i)))P(1 - P(C_i))} \tag{4.1}$$

where C_i represents the i-th CoD and $!C_i$ indicates the compliment of C_i. Over the entire set of possible CoD, $P(C_i|I)$ is normalised in the form:

$$P(C_i|I) = \frac{P(I|C_i)P(C_i)}{\sum_{k=1}^{m} P(I|C_k)P(C_k)} \tag{4.2}$$

The InterVA-4 model developed by Byass et al [6]. provides an initial set of unconditional probabilities for causes of death $C_1, C_2, ..., C_m$ and a matrix of conditional probability $P(I_i|C_i)$ for indicators $I_1, I_2, ..., I_n$ and causes $C_1, C_2, ..., C_m$. A repeated application of the calculation for each $I_1, I_2, ..., I_n$ could be formulated as:

$$P(C_i|I_{1...j}) = \frac{P(I_j|C_i)P(C_i|I_j)}{\sum_{k=1}^{m} P(I_j|C_k)P(C_k|I_j)}, \qquad 1 \leqslant j \leqslant n \tag{4.3}$$

A sequential loop is performed by the InterVA-4 model on all indicators and truncates the probability to 0 if it falls below 0.00001 in the process. In reality, only the recorded indicators are considered by the algorithm in the calculation of the probability. The InterVA-4 measure is therefore the probability of a given cause, conditioned solely by the observed indicators, that is to say:

$$P(C_i|I_{1...j}) = \begin{cases} \frac{P(I_j|C_i)P(C_i|I_{1...j-1})}{\sum_{k=1}^{m} P(I_j|C_k)P(C_k|I_{1...j-1})}, & \text{if } I_j = 1 \\ P(C_i|I_{1...j-1}), & \text{if } I_j = 0 \end{cases} \tag{4.4}$$

The probability for two individuals, therefore, will be conditional on a different number of indicators if the number of indicators reported to have occurred in the two deaths differs. This interpretation typically does not feature prominently in the presentation of results. Instead, a ranking across probabilities within each individual determines the cause classification.

One of the major challenges of using this model is building a matrix of conditional probabilities $P(I|C_i)$ covering all causes of death [6]. The package InterVA-4 adopted the conditional probabilities and unconditional prior probabilities of the causes from the InterVA-4 software which was estimated from a diversity of sources. In particular, the unconditional prior causes incorporate minor changes in response to the level of HIV/AIDS and malaria, which are specified by the user.

The output of InterVA-4 model is a text file with CSV. Table 1 shows the possible output for each death record. The MALPREV and HIVPREV come from the input of the user.

No.	Variable	Description
1	ID	identifier from batch file
2	MALPREV	selected malaria prevalence
3	HIVPREV	selected HIV prevalence
4	PREGSTAT	most likely pregnancy status
5	PREGLIK	likelihood of PREGSTAT
6	PRMAT	likelihood of maternal death
7	INDET	indeterminate outcome
8	CAUSE1	most likely cause
9	LIK1	likelihood of 1st cause
10	CAUSE2	second likely cause
11	LIK2	likelihood of 2nd cause
12	CAUSE3	third likely cause
13	LIK3	likelihood of 3rd cause

Table 1: InterVA-4 Output Variables

Given the fact that the sources code of InterVA-4 method and software proposed by Byass et al. are not readily available (see www.interva.net for further details), this project used the open source function of InterVA-4 implemented for the R community. This function of was designed to take the input of VA data and specified model parameters and output an excel file in .csv format while saving the results in R as well. The call to InterVA-4 function has the following structure in R:

```
1: procedure INTERVA
2:     if filename = path then
3:         InterVA(Input, HIV, Malaria, directory, filename="VA_result",
           output="classic", append=FALSE, replicate=FALSE).
```

With the exception of `HIV` and `Malaria` which are required and given by the user in term of its level/prevalence in the area, the rest of the parameters are optional. The matrix of

the VA dataset is automatically generated by the platform from the underlying database. It should be in the form of a data matrix with each row representing a record of VA data. The matrix should have 246 columns, where the first column is the anonymised ID of the death record and the rest being 245 binary indicators in the specified order predefined in the InterVA-4 model.

4.4 GEOGRAPHIC INFORMATION SYSTEM (GIS) LAYER

OHDSS collects GPS data of all houses in the Demographical Surveillance Area (DSA) during the HDSS baseline data collection and during the routine data collection for the new houses. This system must retrieve geographic data on deaths from the HDSS reference database. To achieve this, the GPS data are automatically extracted and processed from the DSS reference database through a trigger. Thus, given that each death is related to a house, a Google map API layer framework has been used to map the HDSS GPS data extracted and processed as shape files to display on a map, the different causes of death of the deceased grouped by the CoD broad categories.

This component allows the adding or updating of the CoD broad categories (Communicable, Non-communicable and Injuries) and their related CoD generated by the InterVA-4 function. In Figure 4.7 each red icon represents a CoD of communicable diseases as requested by the user. A click on each icon gives the information on the cause of death. Other annotated information can be integrated in future work.

It is up to the user to select the area (formal[1] or informal[2]) and to specify the cause of death category group that the system must display on the online dynamic map. The dynamic use or display of the distribution of CoD through this map is done by selecting the area (formal or informal) and entering the group name of the CoD broad category in the

1 Formal areas are areas develop in presence of government planning processes. In some cases, buildings and neighbourhoods are built legally on agricultural land that is officially assigned for housing and construction

2 While Informal areas concerns the areas where buildings and neighbourhoods are built illegally on agricultural land that is not officially assigned for housing and construction.

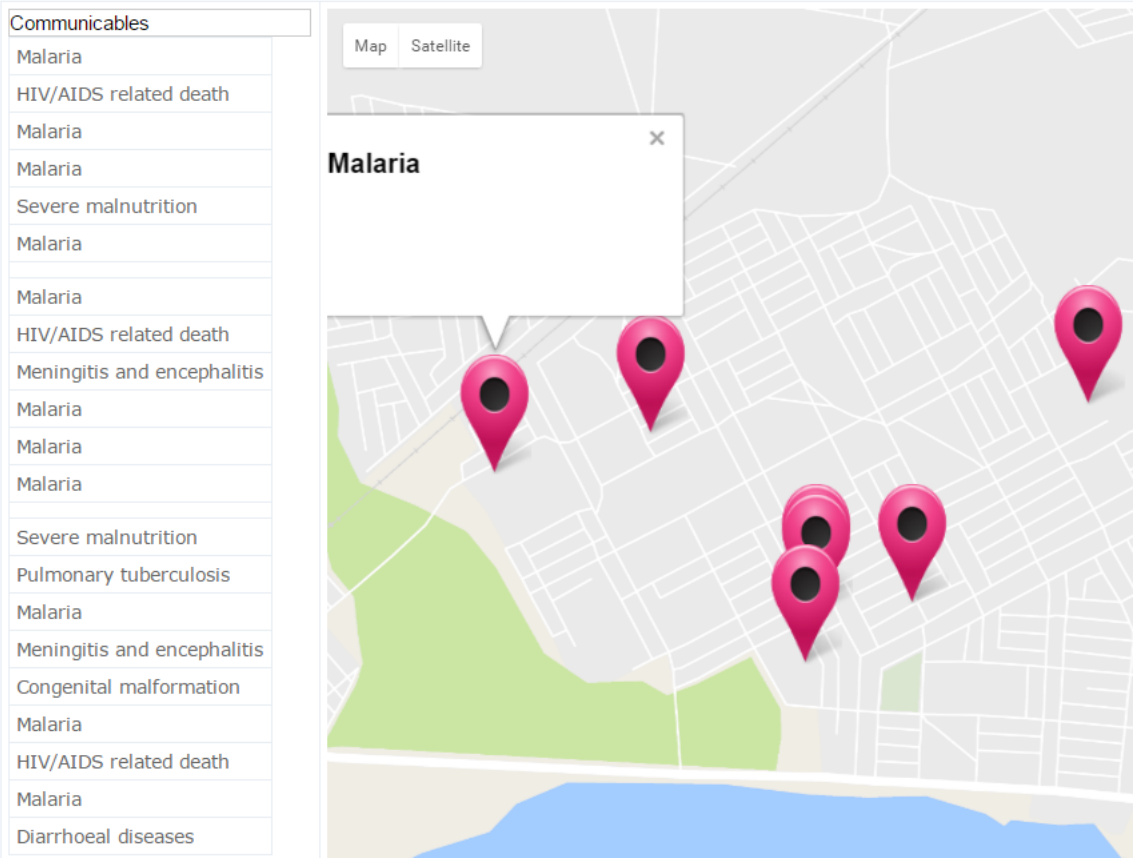

Figure 4.7: Causes of Death Distribution from Communicable Disease in a Formal Area (Tanghin)

reserved text area on the user interface. The area information has been integrated to allow comparison of CoD distribution on maps of between formal and informal settlements within HDSS. As this is a dynamic map interface, the entry of the CoD category group name through the user interface as well as CoDs related retrieval and display through their GPS coordinates from the database are instantaneous. The CoD distribution on these two maps concerns the CoD distribution of one informal area of Ouagadougou Urban HDSS. These types of maps can be used to determine whether the CoD distribution has an impact or effect on the population area or environment.

4.4.1 *Integrating VA Datasets with Google Map API*

The Google Maps API is an application programming interface created by Google to display maps. Among the API widely used in .NET environment for web applications,

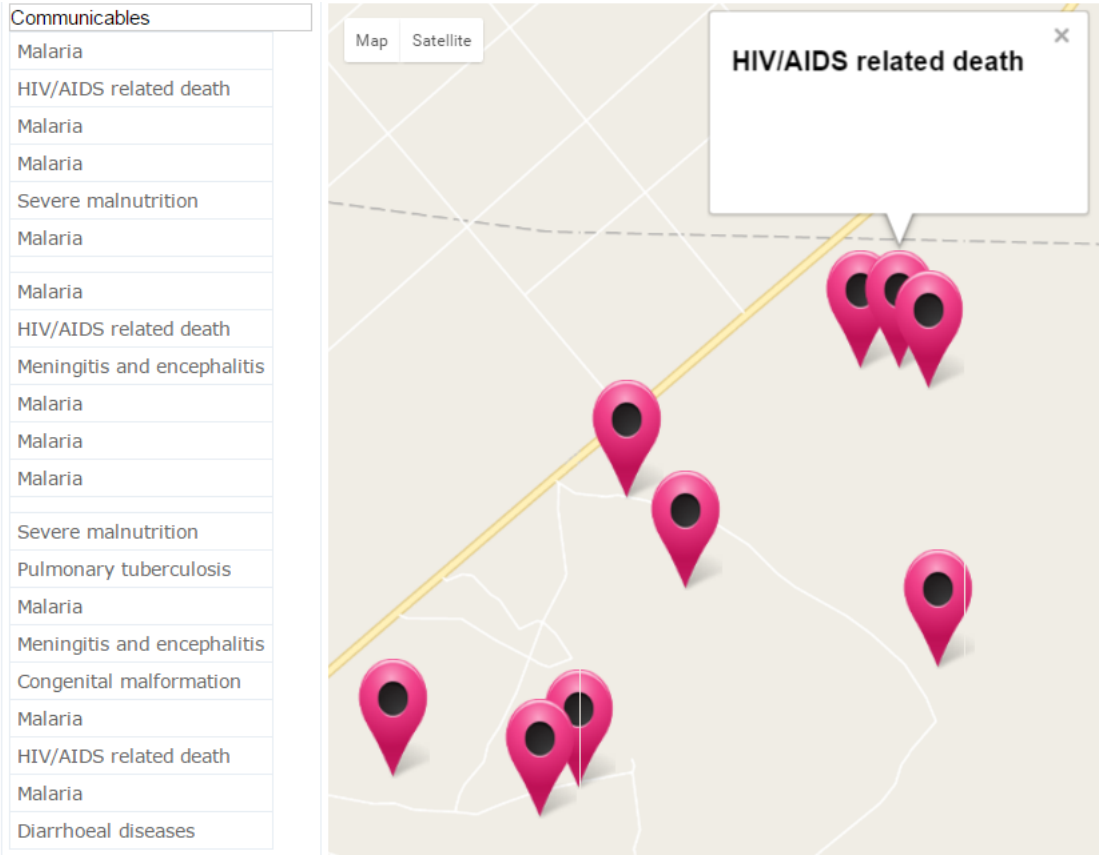

Figure 4.8: Causes of Death Distribution from Communicable Disease in a Informal Area (Nioko2)

this API is suitable for this project because it provides geographic data for maps applications in .NET. We utilised functions such as `google.maps.Map()`, `map.mapTypes.set()`, `map.setMapTypeId()` within this API in conjunction with JavaScript and Active Server Page (ASP.NET) C# to develop the GIS layer. The GPS data associated with the VA datasets is pulled from the database and sent to `google.maps.Map()` and `document.getElementById()` functions to produce a map containing all CoDs.

4.4.2 *Ethical Consideration of Displaying CoD on Maps*

Although GIS is considered as one of the most important technologies in the public health, it has some challenges related to ethical use of personal and health information. However, there has been advances in protection of privacy and confidentiality [16] of information that are being used to address challenges.

In this project, some precautionary measures have been integrated into the platform (See Apendix C)A.1. The first is through the protection against unauthorised use of the platform with the Authentication Use Case described in section 4.1.1. The implementation of this use case obligates each user to be authenticated before access to the module. Furthermore, the identifiers of deceased individuals have been anonymised by using Safe harbour [8] character shifting technic. In addition, an integration of GPS data fields such as roads and street names have been deactivated on the maps to hide HDSS area localisation. The google API used for causes of death distribution on map has settings to hide area and road information, which have been customised to take into account some of these ethical considerations.

The results of this project do not contain any personal identifiers, and the protocol was approved by the University of the Witwatersrand Human Research Ethical Committee (HREC). It could be published in a journal according to the university, the Dodowa and the Ouagadougou HDSSs regulations and standing orders. The protocol of this study was submitted to and approved by the University of the Witwatersrand Faculty of Health Sciences Human Research Ethics Committee under the clearance number, M151029 on 25th of November, 2015.

4.5 LIMITATIONS OF THE VA DATA MANAGEMENT PLATFORM

The following are some limitations of the platform and some challenges encountered during the development:

- Verbal autopsy questionnaires contain multiple skip logics, but our module has not integrated the management of these skip logics. Thus, an improvement of this verbal autopsy automatic form generation is left for future work.

- We experienced challenges with translating the generated form to ODK Collect format. This challenge is due to the complexity of the translation and mapping of

automatic generated XHTML format and the dynamic questionnaire metadata code to an XML format that ODK accepts. This area requires further studies that could propose frameworks that will allow such automatic translation from such dynamic XHTML to XML as well as the validation of the obtained XML file.

- Further work is necessary to improve data security, due to the sensitive nature of the personalised health information that this platform produces and operates on.

5

IMPLEMENTATION AND TESTING

In this chapter, we describe experimental environment and the evaluation of the platform.

5.1 EXPERIMENTAL ENVIRONMENT

The development of this web application was done with Microsoft's Visual Studio ASP.NET Model View Controller (MVC5) tools in a .Net framework 4.6.1 environment. We used the C# programming language and R.NET which uses the R statistical software engine to communicate with ASP.NET application. Since the Ouagadougou HDSS live Database is running on Microsoft SQL SERVER 2014 DBMS, the database component of this application was also implemented in the same DBMS. The platform utilised the three-tier architecture in which the user interface (presentation), functional process logic (*business rules*), computer data storage and data access are developed and maintained as independent modules as illustrated in Figure 5.1.

5.2 DATA SOURCES

Due to time constraint, a practical and production study could not be set up to evaluate this module on actual clinical or health related study. Instead, a sample of completed verbal autopsy questionnaires and data related to deceased individual data (basic information and GPS data) from Ouagadougou (Burkina Faso) and Dodowa (Ghana) HDSSs have been used as entry dataset. Two different sites were used in order to assess if the

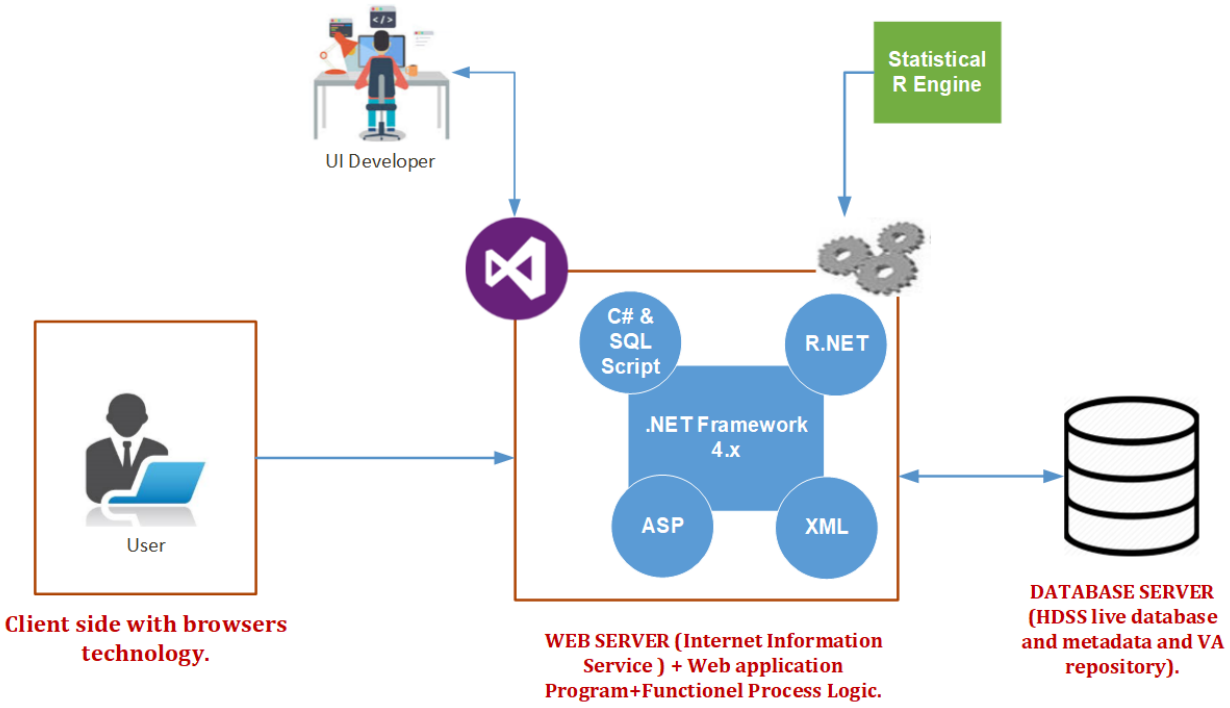

Figure 5.1: Client-Server Software Architecture

system is the best possible fit for the needs of HDSSs and compatible to software of the hardware resources already in place in the two different operational environments.

Twenty-five and 40 verbal autopsy completed questionnaires with data related to deceased individuals were used from Dodowa and Ouagadougou HDSS respectively. The VA questionnaire from Dodowa is an adapted version of the 2012 WHO VA questionnaires and those from Burkina was based on the 2014 WHO verbal autopsy questionnaire. This gives the variation needed for this evaluation. Data related to deceased individuals were in the same format in term of DBMS type and number of variables, which facilitated the data integration process.

5.3 EXPERIMENTAL OUTCOMES

5.3.1 *Experiments Conducted*

The implementation of this system needed the VA dataset from the two HDSS to work. As such, a process was implemented to allow the integrating of the two VA datasets into the verbal autopsy snapshot which is incorporated to the Ouagadougou HDSS reference database.

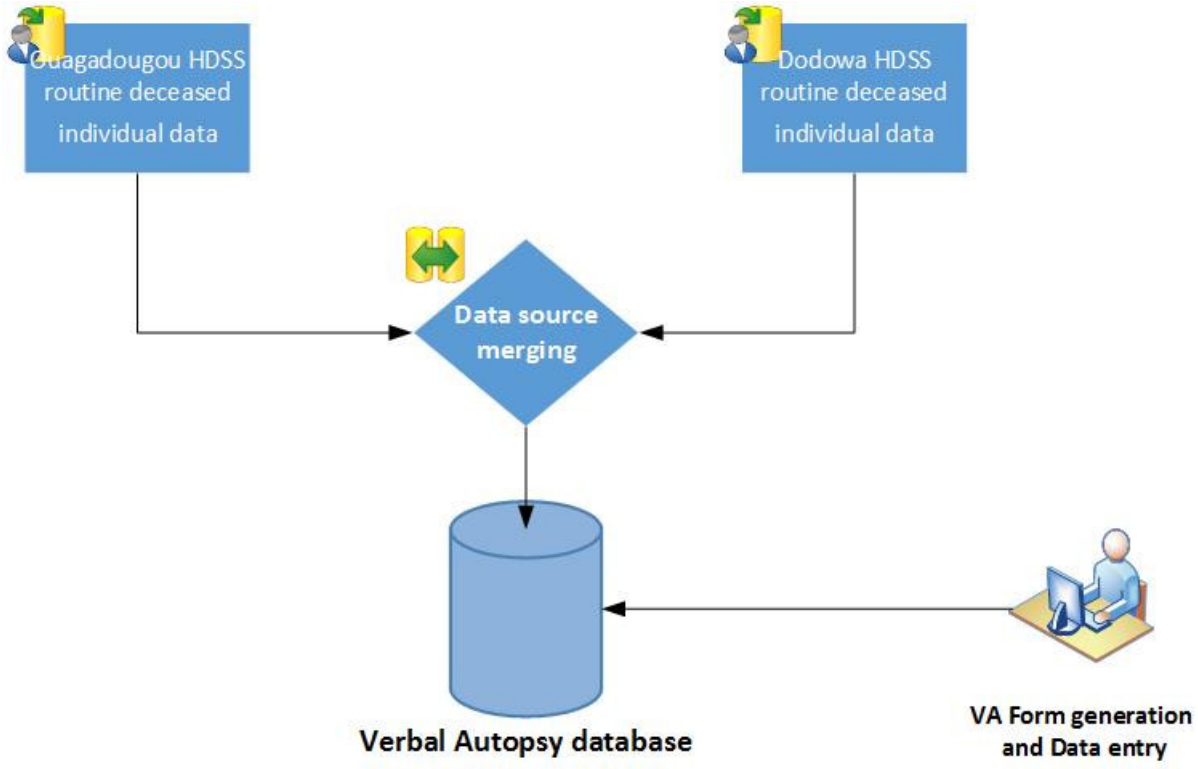

Figure 5.2: Dodowa and Ouagadougou HDSS VA Data Integration Process.

After this data integrating process, the variables codes of each questionnaire was entered into the verbal autopsy metadata form. At the end of this operation, two types of verbal autopsy data entry forms were generated. Fifty VA datasets from all the filled VA questionnaires were entered. Then the verbal autopsy data were converted automatically to the InterVA-4 matrix data format. The matrix data were then processed and cleaned

to ensure that the generated matrix matches with the InterVA-4 format defined by Byass' InterVA-4 software. Throughout this process, the system was checked for data integrity and data quality based on the InterVA-4 Matrix data format requirements. Our tool has a utility that exports the InterVA-4 matrix data for CoD data interpretation with the Byass' InterVA-4 software or another software which accept the InterVA-4 matrix data format.

The obtained InterVA-4 matrix data allowed the automatic CoD interpretation with various parameters selected according to the user's need. This selection of parameters was only provided by the InterVA-4 function of R and not by the Byass' InterVA-4 software, which has less parameters. The selection of the deceased subject's ID in the data entry process allowed the joining of the individual and his or her GPS data automatically. Once the CoD interpretation process was completed, the distribution of CoD could be dynamically displayed on maps.

5.3.2 *Discussions*

The WHO Verbal autopsy questionnaire can be adapted by the research centre to local cultural specifications and practical considerations. However, in the translation of questions responses of the WHO verbal autopsy questionnaire for VA data interpretation, the data format must be consistent with the InterVA-4 matrix data format. While the InterVA-4 algorithm gives satisfactory CoD data interpretation, it is recommended that research centres or researchers directly perform interpretation using InterVA-4 software which is exclusive of the VA management platform. The VA questions of InterVA-4 matrix and the questions of WHO verbal autopsy questionnaire must be identical to facilitate the mapping between these two verbal autopsy questions responses.

With the development of this new system, the upgrading and mapping of interVA-4 Matrix questions to that of WHO verbal autopsy is seamless. The 2016 WHO verbal autopsy is about to be introduced, making this tool relevant. With the rise of other tools such

as Tariff and InSilicoVA for CoD data interpretation, questions of InterVA-4 matrix must be concomitantly upgraded to be consistent with WHO verbal autopsy questionnaire versions to facilitate question mapping and interpretation. In our evaluation, this consistency checking was performed. The mapping was successful irrespective of the version of the WHO VA questionnaire.

There are vital indicators to consider in comparing our platform to the original InterVA-4 software. First, both InterVA-4 of R and Byass' InterVA-4 software work with Common Separate Value (CSV) file. This requires manual work to prepare the csv file and some skills in R scripts or even in statistics. In general, the use of these verbal autopsy data interpretation software needs data science background skills, whereas the developed metadata driven module is accessible to individual having little analytics skills.

Secondly, contrary to Byass' InterVA-4 software and the InterVA-4 of R, this metadata driven module allows verbal autopsy data management through conventional DBMS. The use of R.NET framework in the project allowed the integration of the CoD data interpretation results into the DMBS. The distribution of the CoD on the maps allows researchers to distinguish between the communicable or non-communicable diseases or injuries, and between formal and informal settlements of the two HDSS (Ouagadougou and Dodowa). This distribution of causes of death has shown that over 75% of the CoD related to communicable deceases such as malaria are present in the two types of settlements. However, diseases like diarrhoea and pulmonary tuberculosis are more prevalent in the informal settlements than the formal area. This could be attributed to poor living conditions in these areas. When we exported the InterVA-4 matrix data from this platform to be interpreted by the Byass' InterVA-4 software, it showed slight differences in the CoD probabilities generated by our platform and that of Byass' InterVA-4 software, but within the margin of error. Indeed, our module based on R implements rounding differently from Byass' InterVA-4 software. That occasionally leads to different result. Such irregularities in rounding have only a partial influence on the final results, commonly less than 1% in the calculated likelihood.

6

OPERATIONAL USE CONSIDERATIONS

This chapter highlights the factors and elements that must be taken into consideration in the operational implementation of the platform in health and demographic surveillance systems.

6.1 DIFFERENCES AMONG HDSS SITES

Given that the data model of an HDSS is structured on a defined geographical distribution, each HDSS database design and build are different. In addition, HDSSs use different types of Relational database management systems (RDBMS) to manage their large volumes of data collected over long periods. Some HDSS use proprietary RDBMS system such as Microsoft Access and Microsoft SQL Server (MSSQL), and others use generic systems specifically created for HDSS such as Household Registration System (HDS) or OpenHDS. Some of these DBMSs are empowered to manage and process large volumes of data, and others such as Microsoft Access, Microsoft FoxPro are not advised. Another important point is that some HDSSs, apart from being a member of the INDEPTH Network, are also members of others research networks such as RTS, S-Clinical Trials Partnership and the Alpha Network. That is due to the difference in the needs and the research areas of HDSS researchers. In addition to these differences, there are different methods of creating the identifiers assigned to households and individuals, the naming of the attributes of the various relations (tables) and the operating system environment. In some HDSSs, households IDs are embedded in individuals ones explicitly, while in others this is not the case. For example, in the urban HDSS of Ouagadougou (Burkina Faso), individuals

ID are made up of 19 digits from which the first 12 digits referred to the household ID, but in the rural HDSS of Dodowa (Ghana), the household ID does not form part of the individual's IDs.

All deceased individuals on whom verbal autopsy data must be collected are residents. A resident person in the HDSS is a person who has stayed in the HDSS area for a certain period of time, and the time period varies from site to site. Additionally, some HDSS have three or four rounds of routine data collection per year while others for financial resource reasons have one round of routine data collection in the year.

6.2 OPERATIONAL USE CONSIDERATIONS FOR HDSS SITES

In the development of this platform, an effort has been made to minimise the previously mentioned differences between HDSSs. However, given the complexity of naming the entity variables or attributes of each HDSS, the platform was unable to integrate this difference in order to facilitate the extraction of some data from the death event table in the InterVA-4 matrix. In addition, a view was created in the platform database to give each HDSS the ability to integrate these death data using routine collection as well as renaming the variables to be in line with those of the platform. The database of this platform was designed with Microsoft SQL Server which is the DBMS used by the urban HDSS of Ouagadougou, Dodowa and Agincourt HDSS. The connection to the database under SQL server has been integrated into this platform through a form. This form allows users to enter the database connection settings to allow the platform to access the database. The use of a different DBMS from that used by the platform requires an integration of the drivers of this DBMS. This integration will require a code modification as well as a recompilation of the application. After the connection to the database, the form creation process can then start. A stored SQL procedure integrated in the system must be updated

in order to conform the names of the variables with the new database for the extraction of the relevant VA dataset as well as the GPS data for the GIS layer.

6.2.1 *Recommendations for Deployment and Usage*

The tests were carried out on the data from the Dodowa and Ouagadougou Health and demographic surveillance systems. Considerations to be taken into account for other HDSS include the following:

- Since each type of VA questionnaire is adaptable by each HDSS, it is strongly recommended to retain the codification of the existing variables names (e.g. 3A260, 3A270, etc.) on the WHO VA questionnaires.
- This module was developed to work with MSSQL Server. An extensive design is recommended to handshake with order Database engines.
- The use of this application in a Linux environment requires the use of the Mono module for the hosting of ASP.NET web applications using apache type web services.

7

CONCLUSION AND FUTURE DIRECTIONS

The WHO Verbal Autopsy questionnaires used by several INDEPTH research centres can be adapted to take into account local contextual factors. Thus, in this project we have developed a metadata-driven VA data platform for adequate data management for VA datasets. The platform also accommodates version changes to the WHO VA questionnaire without the intervention of an application developer. The INDEPTH network encourages all HDSS to use good data management practises and relational database management systems to manage VA datasets. This project will help HDSSs to implement these recommendation.

We implemented a web application as a unique tool that help data managers with verbal autopsy data collection, processing as well as interpretation. Data collected using verbal autopsy instruments are entered through an automatically generated verbal autopsy form.

Our platform also incorporates a layer that interprets VA dataset and provides cause-specific mortality fractions of individuals who died in the HDSS area. This automatic interpretation of verbal autopsy data is based on the conditional probability of InterVA-4 function of R incorporated in this layer. The platform provides utilities that allows researchers and data managers to display CoD on maps through a GIS layer. This work is a step towards VA data sharing and cross sites VA investigations and research efforts.

The following activities have been identified for future work:

- The system can be improved by providing annotated information for the CoD distribution on maps. This has been earmarked for future work.
- At the present stage, our module has not integrated the management of skip logics in the automatic CRF generation. Thus, an improvement of this verbal autopsy automatic form generation is left for future work.

- Implementation of a module to facilitate plug and play APIs for the integration of other clinical or cohort studies databases.

A

USE CASE DESCRIPTION AND VERBAL AUTOPSY INSTRUMENT

A.1 USE CASE DESCRIPTION

Use case "Create data entry form" description

Use Case Name: Create data entry form	ID: 1	Importance Level: High
Primary Actor: Administrator	Use Case Type: Essential	
Stakeholders and Interests: User - wants to create his own data entry form for Verbal Autopsy questionnaire		
Brief Description: This use case describes how we create a data entry form for verbal autopsy form as well as changing or cancelling a form metadata.		
Relationships: Include: User Authentication		
Normal Flow of Events: 1. The administrator based on the researcher advice, choose the verbal autopsy questionnaire. 2. The administrator executes the Create data entry form use case.		

Use case "Display causes of death repartition" description;

Use Case Name: Display causes of death repartition	ID: 2	Importance Level: High
Primary Actor: Users	Use Case Type: Essential	
Stakeholders and Interests: User - wants to view on a map the distribution of causes of death from the HDSS area.		
Brief Description: This use case describes the distribution of causes of death by causes of death board categories (Communicable, non-communicable and injuries diseases) from the Ouagadougou HDSS area. This map can help to detect and alert the propagation of some communicable disease		
Relationships: Include: User Authentication		
Normal Flow of Events: 1. The user enter the cause board categories selected. 2. The user executes the Display causes of death repartition use cause.		

Use case "Find probable causes of death" description;

Use Case Name: Find probable causes of death	ID: 3	Importance Level: High
Primary Actor: Users	Use Case Type: Essential	
Stakeholders and Interests: Users - wants to find probable causes of death of the deceased individual.		
Brief Description: This use case describes causes of death data interpretation through InterVA-4 Algorithm.		
Relationships: Include: User Authentication		
Normal Flow of Events: 1. The user decide to analyse verbal autopsy data. 2. The user executes the Find probable causes of death use case.		

Figure 3: Use Case Description

Use case "Entry Verbal Autopsy Death" description;

Use Case Name: Entry Verbal Autopsy Death	ID: 4	Importance Level: High
Primary Actor: Users	Use Case Type: Essential	
Stakeholders and Interests: User - wants to enter the verbal autopsy data into the related database.		
Brief Description: This case describes the verbal autopsy data entry. The data collected with the verbal autopsy questionnaire on the reasons related to the death must be enter into the appropriate database. These data are interpreted using the use case Find probable causes of death.		
Relationships: Include: User Authentication, Create data entry form		
Normal Flow of Events: 1. The user decides to enter the verbal autopsy data collected on the HDSS field. 2. The user executes the Entry Verbal Autopsy Death. 3. It can happen that the entry form has not been created. The user must tell to the administrator to create the entry form.		

Use case "User authentication" description;

Use Case Name: User authentication	ID: 5	Importance Level: High
Primary Actor: Users	Use Case Type: Essential	
Stakeholders and Interests: User - wants to execute a use case.		
Brief Description: This use case describes the user connection to the system when he wants to executes one of the uses cause. The user must enter his authentication information such as username and password.		
Relationships: Include: User Authentication, Create data entry form		
Normal Flow of Events: 1. If the user decides to use one of the use cases. 2. The User authentication use case is popup automatically.		

Figure 4: Use Case Description-Continuation

A.2 VERBAL AUTOPSY INSTRUMENT AND RELATIONAL MODEL

Since the VA questionnaire is too long, below is a web link to it:
http://www.who.int/healthinfo/statistics/verbalautopsystandards/en/

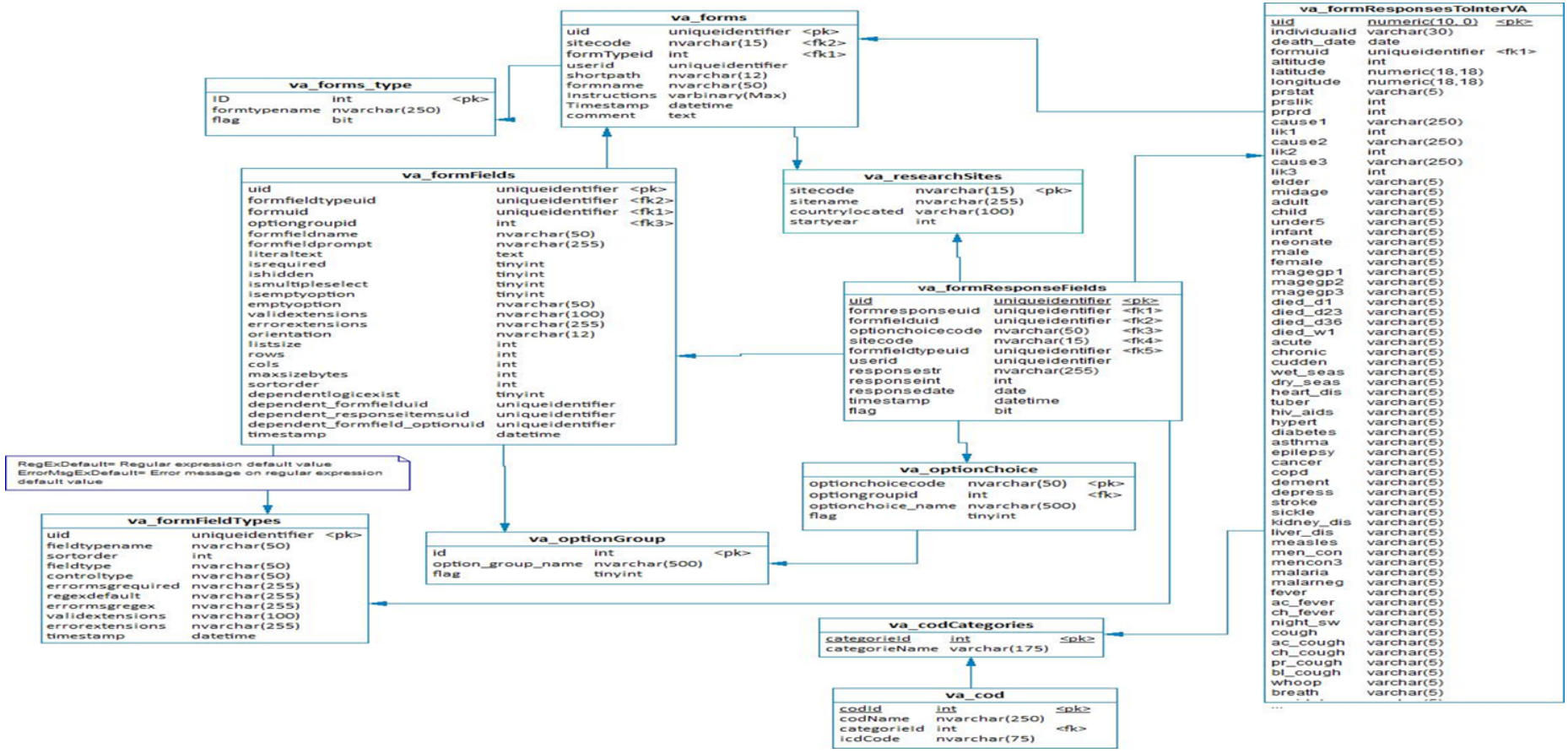

Figure 5: Relational Database Model for the VA Data Platform

B

DATA DICTIONARY

Table	Column or Filed	Description	Data type
va_cod			
	codid	ID of Cause of death table	int
	codname	Cause of death name	nvarchar(250)
	categorieid	Cause of death category (Foreign key)	int
	icdcode	International Classification of Diseases code	nvarchar(75)
va_codcategories			
	categorieid	Primary Key of Cause of death category table	int
	categoriename	Cause of death category description	varchar(250)
va_formFields			
	uid	Primary key of response fields table	uniqueidentifier
	formuid	Foreign key of form table	uniqueidentifier
	formfieldtypeuid	Foreign key of form fileds type table	uniqueidentifier
	optiongroupid	Foreign key of option group table	nvarchar(50)
	formfieldname	form field name	nvarchar(50)
	formfieldprompt	Form field label or prompt	nvarchar(255)
	literaltext	If form field receive Literal text as response	text
	isrequired	If form field response is required	tinyint
	ishidden	If form field is hidden	tinyint
	ismultipleSelect	If form field receive multiple select value	tinyint
	isemptyoption	If form field receive empty response	tinyint
	EmptyOption	Empty option description	nvarchar(50)
	options	Option(Field type)	nvarchar(500)
	validextensions	Valide extension	nvarchar(100)
	errorextensions	Error if mistake in valide extention is required	nvarchar(255)
	orientation	Orientation of the field in the form	nvarchar(12)
	listsize	Size of the liste box	int
	rows	Number of rows	int
	cols	Number of columns	int
	maxsizebytes	Maximum size in byte	int
	sortorder	Sort order key	int
	dependentlogicexist	If the field depends to another field as skip logic	tinyint
	dependent_formfielduid	If skip logic exist, the ID of the field form whose this field depends	uniqueidentifier
	dependent_responseitemsuid	If skip logic exist, the ID of the response fields whose this field depends	uniqueidentifier
	dependent_formfield_optionuid	If skip logic exist, the ID of the form field type whose this field depends	uniqueidentifier
	timestamp	Date of data entry	datetime
va_formFieldTypes			
	uid	Form type ID	uniqueidentifier
	fieldtypename	Form type name	nvarchar(50)
	sortorder	Form field order key	int

	controltype	Field control type	nvarchar(50)
	errormsgrequired	Error message if the the field response is required	nvarchar(255)
	regexdefault	Regular expression of the field if exist	nvarchar(255)
	errormsgregex	Error message of an error in the regular expression of the field	nvarchar(255)
	validextensions	Valid extension to upload if exist	nvarchar(100)
	errorextensions	Error message from the valid extension to upload if exist	nvarchar(255)
	Timestamp		datetime
va_formResponseFields			
	uid	Primary key of response fields table	uniqueidentifier
	useruid	User ID	uniqueidentifier
	formresponseuid	Foreign key of response value table	int
	formfielduid	Foreign key of form fileds table	uniqueidentifier
	optionchoiceid	Foreign key of option choice table	int
	formfieldtype	Field type(Int, date, etc)	nvarchar(250)
	sitecode	Indepth HDSS site code	nvarchar(15)
	responsestr	response in string	nvarchar(255)
	responseint	response in int	int
	responsedate	response in date	date
	individualid	deceased individual ID	nvarchar(250)
	timestamp	Date of data entry	datetime
	flag	Flag for data validity	bit
va_formResponsesToInterva			
	uid		int
	individualid	Identifier of the deceased individual	varchar(250)
	deathid	Anonymised ID of the deceased individual	int
	formuid	Foreign Key of Va_forms table	uniqueidentifier
	dob	Date of Birth of the deceased individual	date
	gender	Gender of the deceased individual	nchar(2)
	death_date	Death date of the deceased individual	date
	altitude	Altitude of the house position of the deceased individual	int
	latitude	Latitude of the house position of the deceased individual	decimal(18, 7)
	longitude	Longitude of the house position of the deceased individual	decimal(18, 7)
	cause1	First most likely cause of death	varchar(250)
	cause2	Second most likely cause of death	varchar(250)
	cause3	Third most likely cause of death	varchar(250)
	elder	Someone aged 65 years or over at death	varchar(5)
	midage	Someoneaged 50 to 64 years or over at death	varchar(5)
	adult	Someoneaged 15 to 49 years or over at death	varchar(5)
	child	Someoneaged 5 to 14 years or over at death	varchar(5)
	under5	Someoneaged 1 to 4 years or over at death	varchar(5)

	infant	Someone aged 1 to 11 months or over at death	varchar(5)
	neonate	Someone aged 28 days or less at death	varchar(5)
	male	Sex of the deceased	varchar(5)
	female	Sex of the deceased	varchar(5)
	magegp1	Women aged 12 to 19 years or over at death	varchar(5)
	magegp2	Womenaged 20 to 34 years or over at death	varchar(5)
	magegp3	Womenaged 35 to 49 years or over at death	varchar(5)
	died_d1	Baby who died within 24 hours of being born	varchar(5)
	died_d23	Babywho died more than 24 hours after birth but within less than 48 hours of being born	varchar(5)
	died_d36	Babywho died more than 48 hours after being born, but within the first week of life	varchar(5)
	died_w1	Babywho died after the first week of life, but within the first month	varchar(5)
	acute	The final illness lasted less than 3 weeks	varchar(5)
	chronic	The final illness lasted 3 weeks or more	varchar(5)
	sudden	The final illness lasted less than a day	varchar(5)
	wet_seas	wet season: Period of death	varchar(5)
	dry_seas	Dry season: Period of death	varchar(5)
	heart_dis	Diagnosis of heart disease	varchar(5)
	tuber	Diagnosis of tuberculosis	varchar(5)
	hiv_aids	Diagnosis of HIV or AIDS	varchar(5)
	hypert	Diagnosis of hypertension, and may include long-term anti- hypertensive medication	varchar(5)
	diabetes	Diagnosis of diabetes made at some kind of medical facility	varchar(5)
	asthma	Diagnosis of asthma made at some kind of medical facility	varchar(5)
	epilepsy	Diagnosis of epilepsy made at some kind of medical facility	varchar(5)
	cancer	Diagnosis of cancer made at some kind of medical facility	varchar(5)
	copd	Diagnosis of chronic obstructive pulmonary disease (COPD) made at some kind of medical facility	varchar(5)
	dement	Diagnosis of dementia made at some kind of medical facility	varchar(5)
	depress	Diagnosis of clinical depression	varchar(5)
	stroke	Diagnosis of stroke made at some kind of medical facility	varchar(5)
	sickle	Diagnosis of sickle cell disease made at some kind of medical facility	varchar(5)
	kidney_dis	Diagnosis of kidney disease made at some kind of medical facility	varchar(5)
	liver_dis	Diagnosis of liver disease made at some kind of medical facility	varchar(5)
	measles	Diagnosis of measles made at some kind of medical facility	varchar(5)
	men_con	Diagnosis of memory loss or confusion made at some kind of medical facility	varchar(5)
	mencon3	Mental confusion lasted for 3 months or more	varchar(5)
	malaria	Positive test for malaria within a week of death	varchar(5)
	malarneg	Negative test for malaria within a week of death	varchar(5)
	fever	Any signs of fever during the final illness	varchar(5)
	ac_fever	Fever lasted less than 2 weeks	varchar(5)
	ch_fever	Fever lasted 2 weeks or more	varchar(5)

	night_sw	Deceased individual experienced heavy night sweats	varchar(5)
	cough	Cough during the final illness	varchar(5)
	ac_cough	Cough lasted less than 3 weeks	varchar(5)
	ch_cough	Cough lasted 3 weeks or more	varchar(5)
	pr_cough	Deceased individual regularly coughed up sputum or mucus	varchar(5)
	bl_cough	Deceased individual coughed up blood	varchar(5)
	whoop	Paroxysms of coughing associated with the characteristic whooping sound of pertussis	varchar(5)
	breath	Any problems of breathing during the final illness	varchar(5)
	rapid_br	An deceased infant or child who was breathing much faster than normal	varchar(5)
	ac_rpbr	Fast breathing lasted for less than 2 weeks	varchar(5)
	ch_rpbr	Fast breathing lasted for 2 weeks or more	varchar(5)
	br_less	Difficulty breathing and was out of breath during the final illness	varchar(5)
	ac_brl	Breathlessness lasted for less than 2 weeks	varchar(5)
	ch_brl	Breathlessness lasted for 2 weeks or more	varchar(5)
	exert_br	Deceased found it particularly difficult to breathe when physically active	varchar(5)
	lying_br	Deceased found it particularly difficult to breathe even when lying down to rest	varchar(5)
	chest_in	Lower chest wall/ribs being pulled in as the child breathed	varchar(5)
	wheeze	Deceased was making wheezing sounds as they breathed	varchar(5)
	ch_pain	Older children or adults who were experiencing severe pain in their chests	varchar(5)
	yellow	Diagnosis of yellow discoloration of the eyes	varchar(5)
	diarr	Diagnosis of diarrhoea	varchar(5)
	ac_diarr	Diarrhoea lasted for less than 2 weeks	varchar(5)
	pe_diarr	Diarrhoea lasted for 2 to 4 weeks, not necessarily every day	varchar(5)
	ch_diarr	Diarrhoea lasted more than 4 weeks, not necessarily every day	varchar(5)
	bl_diarr	Diagnosis of blood in the stoolsduring the final illness	varchar(5)
	vomiting	Any vomiting during the final illness	varchar(5)
	bl_vomit	Diagnosis of vomit contained blood	varchar(5)
	abdom	Diagnosis of any abdominal problem	varchar(5)
	abd_pain	Severe abdominal pain was present during the final illness	varchar(5)
	ac_abdp	Severe abdominal pain lasted for less than 2 weeks	varchar(5)
	ch_abdp	Severe abdominal pain lasted for 2 weeks or more	varchar(5)
	swe_abd	Abnormal distension or protrusion of the abdomen	varchar(5)
	ac_swab	Protuding abdomen lasted for less than 2 weeks	varchar(5)
	ch_swab	Protuding abdomen lasted for 2 weeks or more	varchar(5)
	abd_mass	Diagnostic of any lump inside the abdomen	varchar(5)
	ac_abdm	Diagnostic of a lump inside the abdomen for less than 2 weeks	varchar(5)
	ch_abdm	Diagnostic of a lump inside the abdomen for 2 weeks or more	varchar(5)
	headache	Severe headache was present during the final illness	varchar(5)
	skin	Diagnostic of any skin problem during the final illness	varchar(5)
	skin_les	Dignostic of ulcers, abscess or sores anywhere except on the feet	varchar(5)

	sk_feet	Dignostic of any ulcers, abscess or sores on the feet	varchar(5)
	rash	Dignostic of any skin rash	varchar(5)
	ac_rash	Dignostic of skin rash for less than 1 week	varchar(5)
	ch_rash	Dignostic of skin rash for 1 week or more	varchar(5)
	measrash	Dignostic of measles rash	varchar(5)
	herpes	Dignostic of shingles or herpes zoster	varchar(5)
	stiff_neck	Dignostic of a stiff or painful neck	varchar(5)
	ac_stnk	Dignostic of have a stiff or painful neck for less than 1 week	varchar(5)
	ch_stnk	Dignostic of a stiff or painful neck for 1 week or more	varchar(5)
	coma	Dignostic of a coma or unconsciousness for at least the final 24 hours before death	varchar(5)
	co_ons	Dignostic of the way the deceased became unconscious	varchar(5)
	convul	Dignostic of convulsions or fits during the final illness	varchar(5)
	ac_conv	Dignostic of convulsions last for less than 10 minutes	varchar(5)
	ch_conv	Dignostic of convulsions last for 10 minutes or more	varchar(5)
	unc_con	Dignostic of unconscious immediately after the convulsion	varchar(5)
	urine	Dignostic of any urine problems	varchar(5)
	uri_ret	Dignostic of no urine produced during at least part of the final illness	varchar(5)
	exc_urine	Dignostic of urination was much more frequent than normal during the final illness	varchar(5)
	uri_haem	Dignostic of urine contained blood at any time during the final illness	varchar(5)
	wt_loss	Dignostic of weight loss occurred during the final illness	varchar(5)
	wasting	Dignostic of the extent of the weight loss was sufficiently serious to amount to severe wasting	varchar(5)
	or_cand	Dignostic of oral candidiasis was present during the final illness	varchar(5)
	rigidity	Dignostic of stiffness of the whole body or was unable to open the mouth	varchar(5)
	swell	Dignostic of any lumps	varchar(5)
	swe_oral	Dignostic of any lumps or lesions in the mouth	varchar(5)
	swe_neck	Dignostic of any lumps on the neck	varchar(5)
	swe_armp	Dignostic of any lumps on the armpit	varchar(5)
	swe_breast	Dignostic of an ulcer or swelling in the breast	varchar(5)
	swe_gen	Dignostic of any lumps on the groin	varchar(5)
	swe_oth	Dignostic of swelling (puffiness) of the face	varchar(5)
	swe_legs	Dignostic of both feet swollen	varchar(5)
	anaemia	Dignostic of pale (thinning/lack of blood) or pale palms, eyes or nail beds	varchar(5)
	exc_drink	Dignostic of the person drank a lot more water than usual during the final illness.	varchar(5)
	hair	Dignostic of hair colour change to reddish or yellowish	varchar(5)
	paral_one	Dignostic of paralysis of one side of the body	varchar(5)
	eye_sunk	Dignostic of sunken eyes	varchar(5)
	bl_orif	Dignostic of bleeding from the mouth, nose or anus during the final illness	varchar(5)
	vb_bet	Dignostic of excessive vaginal bleeding in between menstrual periods	varchar(5)
	vb_men	Dignostic of vaginal bleeding stopped naturally during menopause	varchar(5)
	vb_after	Dignostic of vaginal bleeding after menopause	varchar(5)

	diff_sw	Dignostic of difficulty or pain while swallowing liquids	varchar(5)
	not_preg	Dignostic of neither pregnant, nor delivered, within 6 weeks of her death(woman aged 12-50)	varchar(5)
	pregnant	Dignostic of pregnant at the time of death(woman aged 12-50)	varchar(5)
	del_6wks	Woman aged 12-50 years who delivered a baby within 6 weeks of her death	varchar(5)
	pend_6w	Woman aged 12-50 years who have been at an early stage of pregnancy within 6 weeks of her death	varchar(5)
	first_p	Woman aged 12-50 years who have been pregnant for the first time	varchar(5)
	more4	Dignostic of four or more births, including stillbirths, before this pregnancy	varchar(5)
	cs_prev	Dignostic of any previous Caesarean section	varchar(5)
	multip	Dignostic of a woman aged 12-50 years who died while pregnant with twins	varchar(5)
	lab_24	Dignostic of labour for unusually long (more than 24 hours)?	varchar(5)
	died_lab	Woman aged 12-50 years who have been pregnant at the time of her death	varchar(5)
	death_24	Woman aged 12-50 years who have delivered a baby and then died within 24 hours of the delivery	varchar(5)
	baby_al	Woman aged 12-50 years who have delivered a live, healthy baby within 6 weeks of her death	varchar(5)
	breast_fd	Diagnostic of breastfeeding at death	varchar(5)
	del_fac	Dignostic of birth in a health facility	varchar(5)
	del_home	Dignostic of birth at home	varchar(5)
	del_else	Dignostic of birth elsewhere, e.g. on the way to a facility	varchar(5)
	prof_ass	Dignostic of women who receive professional assistance during the delivery	varchar(5)
	del_norm	Dignostic of a normal vaginal delivery	varchar(5)
	del_ass	Dignostic of an assisted delivery, with forceps/vacuum	varchar(5)
	del_cs	Dignostic of delivery by Caesarean section	varchar(5)
	baby_pos	Woman aged 12-50 years who had recently delivered an abnormallly positioned baby	varchar(5)
	mon_early	Dignostic of baby born more than one month early	varchar(5)
	hyster	Dignostic of an operation to remove her uterus shortly before death	varchar(5)
	born_small	Dignostic of baby smaller than normal, weighing under 2.5 kg	varchar(5)
	born_big	Dignostic of baby larger than normal, weighing over 4.5 kg	varchar(5)
	twin	Dignostic of the child part of a multiple birth	varchar(5)
	comdel	Dignostic of child born in a complicated delivery	varchar(5)
	cord	Dignostic of umbilical cord wrapped several times around the neck of the child at birth	varchar(5)
	waters	Dignostic of baby born 24 hours or more after the water broke	varchar(5)
	move_lb	Dignostic of baby stop moving in the womb before labour started	varchar(5)
	cyanosis	Dignostic of baby blue in colour at birth	varchar(5)
	baby_br	Dignostic of baby breathe after birth, even a little	varchar(5)
	born_nobr	Dignostic of baby given assistance to breathe at birth	varchar(5)
	cried	Dignostic of baby cry after birth, even if only a little bit	varchar(5)
	no_life	Dignostic of baby did not cry or breathe, was it born dead	varchar(5)
	mushy	Dignostic of baby born macerated, that is, showed signs of decay	varchar(5)
	fed_d1	Dignostic of baby able to suckle or bottle-feed within first 24 hours after birth	varchar(5)
	st_suck	Dignostic of baby stop suckling of bottle feeding 3 days after birth	varchar(5)
	ab_posit	Dignostic of bottom, feet, arm or hand come into the vagina before its head	varchar(5)

	conv_d1	Dignostic of convulsions starting within the first day of life	varchar(5)
	conv_d2	Dignostic of convulsions starting on the second day or later after birth	varchar(5)
	arch_b	Dignostic of stiff, with the back arched backwards	varchar(5)
	font_hi	Dignostic of bulging or raised fontanelle	varchar(5)
	font_lo	Dignostic of sunken fontanelle	varchar(5)
	unw_d1	Dignostic of unresponsive or unconscious soon after birth	varchar(5)
	unw_d2	Diagnostic of unresponsive or unconscious more than 1 day after birth	varchar(5)
	cold	Diagnostic of cold to the touch before it died	varchar(5)
	umbinf	Diagnostic of redness or discharge from the umbilical cord stump	varchar(5)
	b_yellow	Diagnostic of yellow palms or soles	varchar(5)
	devel	Diagnostic of not growing normally	varchar(5)
	born_malf	Diagnostic of any malformation	varchar(5)
	mlf_bk	Diagnostic of swelling/defect on the back	varchar(5)
	mlf_lh	Diagnostic of a very large head	varchar(5)
	mlf_sh	Diagnostic of a very small head	varchar(5)
	mttv	Diagnostic of tetanus toxoid (TT) vaccine	varchar(5)
	b_norm	Diagnostic of normal vaginal delivery	varchar(5)
	b_assist	Diagnostic of forceps/vacuum	varchar(5)
	b_caes	Diagnostic of Caesarean section	varchar(5)
	b_first	Diagnostic of first pregnancy	varchar(5)
	b_more4	Diagnostic of four or more births, including stillbirths, before this pregnancy	varchar(5)
	b_mbpr	Diagnostic of suffer from high blood pressure	varchar(5)
	b_msmds	Diagnostic of foul smelling vaginal discharge during pregnancy or after delivery	varchar(5)
	b_mcon	Diagnostic of suffer from convulsions during the last 3 months of pregnancy	varchar(5)
	b_mbvi	Diagnostic of suffer from blurred vision during the last 3 months of pregnancy	varchar(5)
	b_mvbl	Diagnostic of vaginal bleeding during the last 3 months of pregnancy but before labour started	varchar(5)
	b_bfac	Diagnostic of baby born in a health facility	varchar(5)
	b_bhome	Diagnostic of baby born at home	varchar(5)
	b_bway	Diagnostic of baby somewhere else, e.g. on the way to a facility	varchar(5)
	b_bprof	Diagnostic of professional assistance during the delivery	varchar(5)
	bpr_preg	Diagnostic of any injury or accident that led to her/his death during pregnancy	varchar(5)
	fit_preg	Woman aged 12-50 years who had blurred vision specifically during her pregnancy	varchar(5)
	vis_bl	Woman aged 12-50 years who had blurred vision specifically during her pregnancy	varchar(5)
	bleed	Woman aged 12-50 years who experienced excessive vaginal bleeding around pregnancy and delivery	varchar(5)
	e_bleed	Diagnostic of vaginal bleeding during the first 6 months of pregnancy	varchar(5)
	s_bleed	Diagnostic of vaginal bleeding during the last 3 months of pregnancy but before labour started	varchar(5)
	d_bleed	Diagnostic of excessive vaginal bleeding during labour	varchar(5)
	p_bleed	Diagnostic of excessive vaginal bleeding after delivering the baby	varchar(5)
	placent_r	Diagnostic of the placenta not completely delivered	varchar(5)
	disch_sm	Diagnostic of foul smelling vaginal discharge during pregnancy or after delivery	varchar(5)

	term_att	Diagnostic of attemption to terminate the pregnancy	varchar(5)
	abort	Diagnostic of pregnancy ended in an abortion (spontaneous or induced)	varchar(5)
	born_early	Diagnostic of a pregnancy which lasted for less than 34 weeks	varchar(5)
	born_3437	Diagnostic of a pregnancy which lasted between 34 and 37 weeks	varchar(5)
	born_38	Diagnostic of pregnancy lasted more than 37 weeks when the baby was born	varchar(5)
	ab_size	Diagnostic of baby of abnormal size	varchar(5)
	injury	Diagnostic of any injury or accident that led to her/his death	varchar(5)
	traffic	Diagnostic of road traffic accident	varchar(5)
	o_trans	Diagnostic of a non-road transport accident	varchar(5)
	fall	Diagnostic of Injured in a fall	varchar(5)
	drown	Diagnostic of death in drowning in water	varchar(5)
	fire	Diagnostic of death from burns?	varchar(5)
	assault	Diagnostic of death from violence/assault	varchar(5)
	venom	Diagnostic of death from any plant/animal/insect bite or sting that led to her/his death	varchar(5)
	force	Diagnostic of death from injury by a force of nature	varchar(5)
	poison	Diagnostic of death from any poisoning	varchar(5)
	inflict	Diagnostic of death from an injury intentionally infection by someone else	varchar(5)
	suicide	Diagnostic of death from suicide	varchar(5)
	alcohol	Diagnostic of regular and sometimes excessive drinker of alcohol up to the time of the final illness	varchar(5)
	smoking	Diagnostic of a regular smoker up to the time of the final illness	varchar(5)
	married	Diagnostic as a married at the time of death	varchar(5)
	vaccin	Diagnostic as adequately vaccinated	varchar(5)
	treat	Diagnostic as receive any treatment for the illness that led to death	varchar(5)
	t_ort	Diagnostic as receive oral rehydration salts	varchar(5)
	t_iv	Diagnostic as receive (or need) intravenous fluids (drip) treatmen	varchar(5)
	blood_tr	Diagnostic as receive (or need) a blood transfusion	varchar(5)
	t_ngt	Diagnostic as receive (or need) treatment/food through a tube passed through the nose	varchar(5)
	antib_i	Diagnostic as receive (or need) injectable antibiotics	varchar(5)
	surgery	Diagnostic as had (or was need) an operation for the illness	varchar(5)
	sur_1m	Diagnostic as have the operation within 1 month before death	varchar(5)
	disch	Diagnostic as discharged from hospital very ill	varchar(5)
	shospf	Diagnostic as travel to a hospital or health facility In the final days before death	varchar(5)
	strans	Diagnostic as use motorised transport to get to the hospital or health facility	varchar(5)
	sadmit	Diagnostic as had problems on arrival to a hospital or health facility	varchar(5)
	streat	Diagnostic as had problems with how they were treated in the hospital or health facility	varchar(5)
	smedic	Diagnostic as having any problems getting medications, or diagnostic tests in the hospital	varchar(5)
	smore2	More than 2 hours to get to the nearest hospital or health facility from the deceased's household	varchar(5)
	sdoubt	Any doubts about whether medical care was needed in the final days before death	varchar(5)
	stradm	Diagnostic as traditional medicine used in the final days before death	varchar(5)
	smobph	Diagnostic as anyone use a telephone or cell phone to call for help in the final days before death	varchar(5)

	scosts	Total costs of care and treatment prohibit other household payments	varchar(5)
	causen	Cause of death description	varchar(250)
	age_category	deceased individual age category	varchar(50)
	area	Death place	varchar(150)
	categorieid	Foreign key of Cause of death category table	int
	timestamp	Date of data entry	datetime
va_forms			
	uid	Primary Key autogenerated	uniqueidentifier
	userid	User ID	uniqueidentifier
	formtypeid	Form field type	int
	shortpath	Short path to access the form	nvarchar(12)
	sitecode	Indepth HDSS site code	nvarchar(15)
	formname	Form name	nvarchar(max)
	instructions	Form use instruction	text
	timestamp	date of the form creation	datetime
va_forms_type			
	formtypeid	ID of forme type table	int
	formtypename	Form type description	nvarchar(250)
	flag	Flag for data validity	bit
va_optionchoice			
	optionchoiceid	Primary key of option choice	int
	optiongroupid	Foreign key of Optiongroup table	nvarchar(50)
	optionchoice_name	Option choice name or description	nvarchar(500)
	flag	Flag for o^ption choice validity	tinyint
va_optionGroup			
	optiongroupid	Form type ID	nvarchar(50)
	option_group_name	Option group name	nvarchar(500)
	flag	Flag for option group validity	tinyint
va_researchsites			
	sitecode	Indepth HDSS site code as Primary Key	nvarchar(15)
	sitename	Indepth HDSS site name	nvarchar(255)
	countrylocated	Indepth HDSS site location(country)	nvarchar(100)
	startyear	Indepth HDSS site year of start	int

C

PLAGIARISM DECLARATION

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG | FACULTY OF HEALTH SCIENCES

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I KOMBASSERE Kouliga (Student number: 1242647) am a student registered for the degree of Masters in Epidemiology (MSc. Research Data Management) in the academic year 2016.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature:  Date: March, 24th 2017

26/04/2015
1

Figure 6: Plagiarism Declaration

D

ETHICS CLEARANCE CERTIFICATE

R14/49 Mr Kombassere Kouliga

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M151029

NAME: (Principal Investigator)	Mr Kombassere Kouliga
DEPARTMENT:	Public Health Ouagadougou HDSS, Institut Superieur des Sciences de la Population, Burkina Faso
PROJECT TITLE:	A Meterdata Driven Module for Managing and Interpreting HDSS Verbal Autopsy Datasets Using Interva-4 Model
DATE CONSIDERED:	30/10/2015
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Gideon Nimako
APPROVED BY:	Professor P Cleaton-Jones, Chairperson, HREC (Medical)
DATE OF APPROVAL:	25/11/2015

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Secretary in Room 10004, 10th floor, Senate House, University.
I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. **I agree to submit a yearly progress report**.

Principal Investigator Signature Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Figure 7: Ethics Clearance Certificate

BIBLIOGRAPHY

[1] Yaw Anokwa, Carl Hartung, Waylon Brunette, Gaetano Borriello, and Adam Lerer. "Open source data collection in the developing world." In: Computer 42.10 (2009).

[2] Amy H Auchincloss and Ana V Diez Roux. "A new tool for epidemiology: the usefulness of dynamic-agent models in understanding place effects on health." In: American journal of epidemiology 168.1 (2008), pp. 1–8.

[3] Frank Baiden, Ayaga Bawah, Sidu Biai, Fred Binka, Ties Boerma, Peter Byass, Daniel Chandramohan, Somnath Chatterji, Cyril Engmann, Dieltiens Greet, et al. "Setting international standards for verbal autopsy." In: Bulletin of the World Health Organization 85.8 (2007), pp. 570–571.

[4] Justus Benzler, Kobus Herbst, and Bruce MacLeod. A data model for demographic surveillance syst 1998.

[5] Peter Byass, Dao Lan Huong, and Hoang Van Minh. "A probabilistic approach to interpreting verbal autopsies: methodology and preliminary validation in Vietnam." In: Scandinavian Journal of Public Health 31.62 suppl (2003), pp. 32–37.

[6] Peter Byass, Daniel Chandramohan, Samuel J Clark, Lucia D'Ambruoso, Edward Fottrell, Wendy J Graham, Abraham J Herbst, Abraham Hodgson, Sennen Hounton, Kathleen Kahn, et al. "Strengthening standardised interpretation of verbal autopsy data: the new InterVA-4 tool." In: Global health action 5 (2012).

[7] Nikita Desai, Lukasz Aleksandrowicz, Pierre Miasnikof, Ying Lu, Jordana Leitao, Peter Byass, Stephen Tollman, Paul Mee, Dewan Alam, Suresh Kumar Rathi, et al. "Performance of four computer-coded verbal autopsy methods for cause of death assignment compared with physician coding on 24,000 deaths in low-and middle-income countries." In: BMC medicine 12.1 (2014), p. 20.

[8] Khaled El Emam. Guide to the de-identification of personal health information. CRC Press, 2013.

[9] Abraham D Flaxman, Alireza Vahdatpour, Sean Green, Spencer L James, and Christopher JL Murray. "Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards." In: Population health metrics 9.1 (2011), p. 29.

[10] Edward Fottrell and Peter Byass. "Verbal autopsy: methods in transition." In: Epidemiologic reviews (2010), mxq003.

[11] Rebecca Guenther and Jacqueline Radebaugh. "Understanding metadata." In: National Information (2004).

[12] Tobias Homan, Aurelio Pasquale, Ibrahim Kiche, Kelvin Onoka, Alexandra Hiscox, Collins Mweresa, Wolfgang R Mukabana, Willem Takken, and Nicolas Maire. "Innovative tools and OpenHDS for health and demographic surveillance on Rusinga Island, Kenya." In: BMC research notes 8.1 (2015), p. 397.

[13] Spencer L James, Abraham D Flaxman, and Christopher JL Murray. "Performance of the Tariff Method: validation of a simple additive algorithm for analysis of verbal autopsies." In: Population Health Metrics 9.1 (2011), p. 31.

[14] Gary King, Ying Lu, and Kenji Shibuya. "Designing verbal autopsy studies." In: Population Health Metrics 8.1 (2010), p. 19.

[15] Jordana Leitao, Daniel Chandramohan, Peter Byass, Robert Jakob, Kanitta Bundhamcharoen, Chanpen Choprapawon, Don De Savigny, Edward Fottrell, Elizabeth França, Frederik Frøen, et al. "Revising the WHO verbal autopsy instrument to facilitate routine cause-of-death monitoring." In: Global health action 6 (2013).

[16] Zehang Richard Li, Tyler H McCormick, and Samuel J Clark. InterVA4: An R package to analyze ve 2014.

[17] Zehang Richard Li, Tyler H McCormick, and Samuel J Clark. InterVA4: An R package to analyze ve 2014.

[18] Lene Mikkelsen, David E Phillips, Carla AbouZahr, Philip W Setel, Don De Savigny, Rafael Lozano, and Alan D Lopez. "A global assessment of civil registration and vital statistics systems: monitoring data quality and progress." In: The Lancet 386.10001 (2015), pp. 1395–1406.

[19] Christopher JL Murray, Rafael Lozano, Abraham D Flaxman, Peter Serina, David Phillips, Andrea Stewart, Spencer L James, Alireza Vahdatpour, Charles Atkinson, Michael K Freeman, et al. "Using verbal autopsy to measure causes of death: the comparative performance of existing methods." In: BMC medicine 12.1 (2014), p. 5.

[20] INDEPTH Network. Population and health in developing countries: Population, health, and surviv International Development Research Centre, 2002.

[21] INDEPTH Network. "INDEPTH standardized verbal autopsy questionnaire." In: Available from: www. indepth-network. org/core_documents/indepthtools. htm (2003).

[22] INDEPTH Network. "INDEPTH resource kit for demographic surveillance systems." In: Accra: INDEPTH Network (2013).

[23] World Health Organization et al. "Civil registration: why counting births and deaths is important." In: World Health Organization (2014).

[24] Dirk Pfeiffer, Timothy P Robinson, Mark Stevenson, Kim B Stevens, David J Rogers, Archie CA Clements, et al. Spatial analysis in epidemiology. Vol. 557976890. Oxford University Press New York, 2008.

[25] Mgr Vojtěch Přehnal. "A Metadata-Driven Approach to Relational Database Management." PhD thesis. Masarykova univerzita, Fakulta informatiky, 2012.

[26] Osman Sankoh and Peter Byass. The INDEPTH Network: filling vital gaps in global epidemiology. 2012.

[27] James K Tamgno, Roger M Faye, and Claude Lishou. "Verbal autopsies, mobile data collection for monitoring and warning causes of deaths." In: Advanced Communication Technology IEEE. 2013, pp. 495–501.

[28] Craig D Weissman and Steve Bobrowski. "The design of the force.com multitenant internet application development platform." In: SIGMOD Conference. 2009, pp. 889–896.

[29] World Health Organization (WHO). ICD-10 Revision (Internet). 2015. URL: http://www.apa.org/international/outreach/who-icd-revision.aspx.

[30] Yazoume Ye, Marilyn Wamukoya, Alex Ezeh, Jacques BO Emina, and Osman Sankoh. "Health and demographic surveillance systems: a step towards full civil registration and vital statistics system in sub-Sahara Africa?" In: BMC public health 12.1 (2012), p. 741.

[31] Carla Abou Zahr. Verbal autopsy standards: ascertaining and attributing cause of death. World Health Organization, 2007.