A Metadata Driven Module for Managing and Interpreting HDSS Verbal Autopsy Datasets using Interva-4 Model

KOULIGA KOMBASSERE (✉ kkombassere@issp.bf)

Institut Supérieur des Sciences de la Population <https://orcid.org/0000-0001-8301-8676>

Gideon Nimako

University of the Witwatersrand

Irma Mare

University of the Witwatersrand

Technical advance

Keywords: Research Database management, Metadata-driven, Data Management, Geographical Information System, Database, Statistical data, Software tools, Data interpretation

Posted Date: June 11th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-30889/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

RESEARCH

A Metadata Driven Module for Managing and Interpreting HDSS Verbal Autopsy Datasets using Interva-4 Model

Kouliga Kombassere^{1*}, Gideon Nimako²
and Irma Mare³

*Correspondence:

kkombassere@issp.bf

¹Universite Joseph KI-ZERBO,
Institut Supérieur des Sciences de
la Population, Burkina Faso,
Avenue Charles de Gaulle,
Ouagadougou, BF

Full list of author information is
available at the end of the article

Abstract

Background: The World Health Organisation provides a standardised survey-based questionnaire for collecting cause of death data. This standardised tool is undergoing iterative changes almost every 3 to 5 years resulting in the redesign of in-use re-adapted questionnaires and their database schemes. Given the size of this questionnaire, its redesign process requiring a lot of time and resources does not allow research centres to update their questionnaire. In addition, the heaviness and the expensive cost of the Physician Certified Verbal Autopsy method used for collected data interpretation, led the emergence of new methods with which data are usually managed in ad hoc fashion by using spreadsheets and Comma Separated Value files. Therefore, these tools not allow preservation of the contextual metadata and also not support recovery and building relationship among data object. While the absence of data object relationships does not facilitate the use of relational database management systems and data preservation over time in longitudinal studies contexts such Health and Demographic surveillance systems.

Results: This research used Microsoft Visual studio based on model-driven and metadata architectures associated with R.NET Package, R InterVA function, Google Maps API, eXtensible Mark-up Language and Microsoft SQL Server 2012 to develop a Verbal Autopsy data management platform. This platform assists INDEPTH Network HDSS fields sites to quickly follow the iterative changes of WHO questionnaire through questionnaire generation, data collection and entry, and a mapping layer that translates verbal autopsy CRF to ODK XML data dictionary enabling cause of death data collection in offline mode using handheld devices. In addition, being a R InterVA function aided tool for the interpretation of cause of death data, this web application has an interface for visualising cause of death patterns using Google Maps API.

Conclusions: Verbal autopsy data management interpretation over time in the longitudinal studies context such Health and Demographic Surveillance System is feasible. Thereby, possibilities given by metadata-driven to build reliable software architecture for verbal autopsy data collection, interpretation and cause of death patterns visualisation and particularly compliance with regulatory relational database management requirements are achievable.

Keywords: Research Database management; Metadata-driven; Data Management; Geographical Information System; Database; Statistical data; Software tools; Data interpretation

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, most of such challenges are in developing countries[1]. This shows that very few developing countries have functioning cause of death information systems that they can draw on to guide policies for health programs. The lack of reliable data on the levels and causes of death 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. To assist policy makers and international organizations in increasing mortality registration, researchers and experts of the World Health Organization (WHO) have developed three types of standardized and harmonized questionnaires to collect cause-specific mortality data specifically in developing countries[2]. The purpose of these three types of questionnaires (see appendix 8.1) 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). Information on symptoms and the 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 causes of death is called Verbal Autopsy (VA). From 2007 to 2016, under the leadership of WHO, causes of death collected with these various versions of questionnaires are listed and codified based on International Classification of Diseases (ICD-10)[3]. Although these various versions exist, VA questionnaires could be modified and adapted to the context and the local language[2]. Verbal Autopsy data collected are interpreted with a method called Physician Certified VA (PCVA)[4]. However, given the challenges related to its high cost, slowness, non-repeatability and reproducibility, others computers aided methods such as Interpreting 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 to facilitate causes of death interpretation[5]. Although InterVA-4 gives satisfying automated interpretation in most Health and Demographic Surveillance System sites, its use has some challenges. The main challenge, according to World Health Organization, 2007 is the collection of VA dataset with varying versions (2007, 2012, 2014 and 2016) of the WHO VA instrument and difficulties in analysis of the VA dataset sampled over several years for multiple HDSS sites due to the inconsistencies in the data formats and structures. In addition, VA datasets are usually managed in ad hoc fashion by using spreadsheets and Comma Separated Value (CSV) files which do not meet the recommendation of INDEPTH Network for the HDSS data management[6]. For such vital public health registration systems and sorely having system's that allows for longitudinal data analysis is crucial for policy design and interventions. Although there have been efforts to automate VA data interpretation, there is no data management platform for conventional VA datasets as well as Good Clinical Practice (GCP) for VA. In addition, there is overly no tool to dynamically visualize and display the distribution of causes of death for a HDSS, serving as an early epidemiological monitoring and warring tool. Also, it is imperative that these

research centres utilize an adequate data management platform to produce vital data which may be linked to similar data from other HDSSs or hospital recording systems to produce more representative insight and to reduce the lack of vital statistics, health indicators and medical certification.

Background and Related Work

Verbal Autopsy (VA)

According to the WHO [7] most deaths without registration or certification occur in developing countries and are usually attributed to the inaccessibility of health facilities and limited for certain cultural considerations [8]. This has resulted in the emergence of verbal autopsy, which attempt to determine causes of death for previously undocumented deceased persons, thereby allowing scientists to analyse disease patterns and direct public health policy decisions. The process consists of a trained interviewer using the full or an adapted WHO standardized questionnaire to collect information about the signs and symptoms of events of the deceased person from his next of kin or other caregivers [9]. Since 2007, Verbal Autopsy dataset collected has been analysed by health professionals who manually assign a probable cause of death. This analysis method (known as the Physician Certified VA (PCVA) utilizes the International Classification of Diseases (ICD) diseases codification to assign the closest probable cause of death [2]. However, the lack of reproducibility, slowness and high costs associated with this method have led to the computerization of the VA analysis process [10]. Thus, the WHO verbal autopsy instrument has undergone four revisions (2007, 2012, 2014, and 2016) 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, such as the adaptation of questions to the local context and the integration. The computerization of the verbal autopsy data collection, interpretation and management process is still challenge for most death facilities and research institution[11].

VA Computerized Methods

Among Computerized Coding of VA (CCVA) methods there are several algorithmic methods which follow a set of predefined diagnostic criteria giving binary outcomes (yes or no) for a single cause of death. Other methods employ data-driven machine-learning and probabilistic techniques that can give the probability of multiple causes of death from a death. These methods are InterVA-4 [10], King-Lu (KL) direct algorithm, Tariff (SmartVA- Analyse), Random Forest (RF), InSilicoVA Algorithm, Artificial Neural Network (ANN) and the Simplified Symptom Pattern (SSP), etc. Based on the International Classification of Diseases list as reference, these computerized methods were developed with the goal of facilitating causes of death interpretations[5, 12]. Among these tools, WHO recognizes the Tariff, InterVA and recently InSilicoVA algorithm as VA data analytical tools [13]. HDSS sites and research centres such as Institute for Health Metrics and Evaluation (IHME) are currently using it [14, 15]. In the same way, Agincourt HDSS has started computerizing the data collection of causes of death using their adapted VA questionnaire on mobile ad handled devices [16] attempted to simplify real time data capture. This computerized tool, called Mobile InterVA (MIVA) does not permit the dynamic creation of VA questionnaire forms which can accommodate frequent version

changes and the local area context [16]. It also has not any visualization utility to display causes of death repartition on dynamic maps and it does not use the data model suggested by the INDEPTH Network. For the InterVA-4, InSilicoVA and Tariff documentations, the input data must be in Comma Separated Values (CSV) format. As a result, most research centres store VA data in CSV format. This does not allow preservation of the contextual metadata and also does not support recovery and building relationship among data object. For this reason, longitudinal study with a dynamic cohort, such as HDSS, the Reference Data Model (RDM)[17], is recommended by INDEPTH Network. The consideration of this data model facilitates the manipulation, management and preservation of data in a defined area over time. The InterVA of R package kernel has been implemented based on the model underlying InterVA-4 as published by Byass et al, 2012., and uses the Bayes theorem to determine the conditional probability of each given cause of death as a function of a set of signs, symptoms and circumstances listed in the VA questionnaire as a set of binary indicators representing whether the event occurs or not. Thus, the conditional probability for each cause of death could be calculated as follows:

$$P(C_i|I) = \frac{P(I|C_i)P(C_i)}{P(I|C_i)P(C_i) + P(I|!C_i)P(!C_i)} \quad (1)$$

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 normalized in the form:

$$P(C_i|I) = \frac{P(I|C_i)P(C_i)}{\sum_{k=1}^m P(I|C_k)P(C_k)} \quad (2)$$

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

$$P(C_i|I_{1...j}) = \frac{P(I_j|C_i)P(C_i|I_{1...j-1})}{\sum_{k=1}^m P(I_j|C_k)P(C_k|I_{1...j-1})} \quad (3)$$

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:

$$P(C_i|I_{1...j}) = \begin{cases} \frac{P(I_j|C_i)P(C_i|I_{1...j-1})}{\sum_{k=1}^m P(I_j|C_k)P(C_k|I_{1...j-1})}, & \text{if } (I_j = 1) \\ P(C_i|I_{1...j-1}), & \text{if } (I_j = 0) \end{cases} \quad (4)$$

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 major challenge of using this model is building a matrix of conditional probabilities $P(I=C_i)$ covering all causes of death. [18]. 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. Given the fact that the sources code of InterVA-4 method and software proposed by Byass et al., 2012 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 was designed to take as input VA data and some predefined parameters and output a spreadsheet file in Comma-Separated Values(csv) format while saving the results in R as well.

Related Works on Public HDSS Data Systems

Prior to the development of Open Health and Demographic System (OpenHDS), common classes of the HDSS data reference model were not taken into account in HDSS database design [19]. These common data classes 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 and versioning. As a result, the goal of creating a standardized model for data sharing within the INDEPTH Network and helping research centres to develop accurate and reusable software is not well achieved [17]. 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 development and the use of OpenHDS which is a metadata driven model based on the static INDEPTH reference data model[19]. 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 which allow the cause of death visualization. Hence the importance of this work, 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 show that there is a paucity of tools allowing data capture form generation for verbal autopsy data capture in the INDEPTH Network research centres. Especially for Burkina Faso and Ghana, ECOWAS member countries, where health information and surveillance systems fail to detect and enable rapid decision-making in the face of recurrent outbreaks of epidemics such as Cholera, Ebola causing death and economic loss in the region. In such a context 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. The tool enables data capture form generation, verbal autopsy data interpretation using InterVA-4, the visualization of the location of the diseases and distribution of CoD. This tool is an effort that allows researchers and policy makers to plan and design appropriate programmes to provide interventions for communicable and non-communicable diseases in resource limited communities. This development of applications is also a great advantage to reduce tools used in the collection and interpretation of data from verbal autopsies.

Results

Metadata Driven Software Development

In recent years, there has been increased demand for applications that allow modification of the underlying data structures with no programming effort. This new demand requires new systems development techniques for data applications that enable schema alteration in real-time [14]. A metadata-driven approach to relational database design has been used as a response to this challenge and to reduce the time and the effort developers spend on programming during system updates and/or upgrades [20].

Metadata is defined by the National Information Standards Organization (NISO) as “structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource” [21]. Most metadata schemes or syntaxes are expressed in a number of different markups or programming languages such as Hypertext Markup Language (HTML), eXtensible Markup Language (XML), etc., each of which require a specific syntax to structure its contents. In database management systems (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. Metadata has contributed to the health research software development through applications such as Research Electronic Data Capture (REDCap), Census and Survey Processing System (CSPro), Open Data Kit (ODK) etc. which are dynamic content generation applications.

Metadata-driven applications are vital in scenarios that require dynamic or automatic content generation such as Verbal Autopsy questionnaire Case Report Forms (CRF), based on metadata stored in a database, including variable name, variable type, question prompt etc. Since metadata is also data, it is easier to manage forms creation elements stored as data in the application relational database than to manage source code. Such metadata-driven development model is utilized to enable the re-adaptability of the verbal autopsy questionnaires in HDSS research centres and to help HDSS Data managers to make the changes of verbal autopsy questionnaires CRFs by controlling all the rich behaviour of form fields, including formatting, validation, visibility, user interface(UI) type such as drop-down vs text field vs radio buttons, etc.. From the literature, the approaches developed to generate on the fly graphical user interfaces and database schema can be summarized in two main approaches: The approach where graphical user interfaces are generated from metadata stored within a database; the last approach where metadata information is captured using the UI to automatically create and/or update database table information.

The first approach is used in the verbal autopsy questionnaires CRFs generation according to the VA Questionnaire version in Ouagadougou urban HDSS. This approach allows the end user to automatically generate and dynamically manage UIs from the database without any script or coding. Its benefits are threefold: 1) there is an absence of a recompilation of the application code and the redeployment of components on the presentation layer as the customization 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 no skills in application coding 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. The repository storage could be a Relational Database Management System such as Microsoft SQL Server or any other data store format like XML files. The figure 1 shows the metadata driven process schema and overview of the design process. The designer develops a graphical UI through metadata tags that are stored in the relational database, which allow the end user to dynamically create and automatically generate a UI by using web APIs. The main data used to create and generate the UIs as well as the data entered with the generated interface are stored in a single database. And the figure 2 shows the mapping 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 first table stores the data for each UI project created by the end user. Several objects can be related to a project. Each row of these objects in the object or the content table can contain multiple controls such as text fields, combo boxes, radio buttons etc. This diagram provides an overview of what might be a relational database schema for setting up a metadata driven application for automatic UIs generation.

Metadata Driven Platform for Verbal Autopsy

The work is based on the implementation of the metadata design frameworks to replace the ad-hoc tools used by most HDSS sites for the verbal autopsy data capture and interpretation. All the processes of the data collection, the data interpretation and management are done using a unique platform:

- 1 *Create Data Entry Form* allows any user to create data entry forms 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* 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 *Find Probable Causes of Death*.
- 5 *User Authentication* explains the user connection to the system during the use or the running of one of these use cases.

Figure 3 shows the use cases diagram model that the system needs to perform. The descriptions on these use cases meanings are given by figure 10. And the figure 11 shows the relational database model built based on these uses cases for the VA Platform using InterVA-4 model.

Based on these use case actions and the data model, five main components have been implemented as the system functionalities. The first component deals with VA data collection instrument generation and the second allows VA data entry while the third (ODK data dictionary) component handles the translation of the CRFs forms to ODK formats. Figure 4 shows the architecture on which the verbal autopsy system was developed. The fourth component (green rectangles) manages the interpretation of the verbal autopsy entered data using InterVA-4 algorithm which are based on Bayes' theorem. In R package, the InterVA function from this algorithm has the following structure [22]:

```
InterVA (Input, HIV, Malaria, directory = NULL, filename = "VA_result",
output = "classic", append = FALSE, replicate = FALSE)
```

Figure 5 shows the form for verbal autopsy questions metadata entry form extract where questions codes are entered by end user to generate the verbal autopsy data entry form. Each code like 3A260, 3A270, 3A280, etc 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 organized 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 it through a form dedicated to the entry of metadata. A RenderHtml() function takes as arguments or parameters 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.

The matrix of the VA dataset with 246 variables is automatically generated from the underlying database where each row representing a record of VA data. The first column is the anonymised ID of the deceased individual ID into the HDSS Database and the rest in the specified order predefined in the InterVA-4 model. As verbal autopsy variables are not identical to the variables of this matrix, the InterVA-4 User Guide (version 4.RC1 2012-08-14) was used as a reference to automatically map these two variable sets. The InterVA-4 matrix is also composed of questions from the three types of VA questionnaire with much more precision on the types of responses. All the variables in the matrix are closely related to all questions emanating from the three types of questionnaire. This has been taken into account in the codification of the VA questions, in order to facilitate the link between questions of verbal autopsy and those of InterVA-4. As a result, the three types of questionnaire have been coded in a single, coherent in order to save time needed for data extraction.

Interpretation of Verbal Autopsy Data

The VA data interpretation precedes its cleaning. This cleaning processes are managed by stored procedures based on predefined logic. Once the data is cleaned, its interpretation begins with InterVA-4 function of R which takes as input the matrix data and other specified model parameters. The InterVA-4 function which is written in R has been converted to allow its use with Microsoft .NET (See appendix ??). This function is executed by the system through the extracted and cleaned datasets to obtain the CoD for each deceased individual. Figure 6 shows this result except values of variables ID, MALPREV, HIVPREV that are automatically generated by the function can be exported into Excel 2007 or OpenOffice Calc spreadsheet and earlier versions. The content of the ID variable is an anonymised individual ID from the HDSS routine database. The values of variables MALPREV, HIVPREV represent the prevalence of Malaria and respectively HIV given as parameters during the execution of the R InterVA function. The figure 7 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 interpretation of VA data from Ouagadougou Urban HDSS using the function of R. The researchers' tool can also indicate the probability of people to die of a specific disease such as Malaria, HIV/AIDS, Severe Malnutrition in given area, etc. And the figure 8 shows that the probabilities for Malaria and acute respiration infection in the HDSS area are higher than the rest of the causes of death. All the statistics accompanied by charts are automatically generated from the VA cleaned data.

Any HDSS collects Global Positioning System (GPS) data of all houses in the Demographical Surveillance Area as Ouagadougou HDSS during the baseline and routine data collection. These GPS data are automatically extracted and processed from the HDSS database through a trigger using a Google map API layer framework as shape files to display different causes of death grouped by the CoD broad categories on a dynamic map. Figure 9 shows causes of death distribution of communicable disease in an Ouagadougou HDSS formal area, automatically obtained through the verbal Autopsy platform after data entry.

The causes of death have been classified based on the broad categories (Communicable, Non-communicable and Injuries) of decease. Based on this classification, each red icon represents a CoD using the InterVA function and GIS functionalities. These types of cards can be used, in the collection of causes of death, as warning or monitoring systems for certain known communicable and non-communicable diseases such as Cholera, Zika, Ebola and unknown diseases that could have incalculable human and economic consequences.

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. • The authors 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 to allow such automatic translation from 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 personalized health information that this platform produces and operates on.

Data Sources Used

Twenty-five and forty verbal autopsy completed questionnaires with data related to deceased individuals were used from Dodowa Rural Health and Demographic Surveillance System (DRHDSS) and Ouagadougou Urban Health and Demographic Surveillance System (ORHDSS) respectively. The VA questionnaire from DRHDSS is an adapted version of the 2012 WHO VA questionnaires and those from ORHDSS 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;

Discussions

The implementation of the system needed the VA dataset from the two HDSSs. 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. After the data integrating process, the variables codes of each questionnaire were entered into the verbal autopsy metadata form. At the end of this operation, two types of verbal autopsy data entry forms were generated. Then, the verbal autopsy data was converted automatically to the InterVA matrix data format. The matrix data were then processed and cleaned to ensure that the generated matrix matches with the InterVA format defined by Byass' InterVA-4 software. Figure 12 shows this process in which the system was checked for data integrity and data quality based on the InterVA Matrix data format requirements. The 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 matrix data format. The obtained InterVA 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 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 data and his GPS data automatically. Once the CoD interpretation process was completed, the distribution of CoD could be dynamically displayed on maps. 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, the data format must be consistent with the InterVA matrix data format. While the InterVA algorithm gives satisfactory of 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 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 matrix must be concomitantly upgraded to be consistent with WHO verbal autopsy questionnaire versions to facilitate question mapping and interpretation. In the evaluation process, 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 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 VA data interpretation software needs data science background skills, whereas the developed metadata driven module is accessible to individuals having little analytics skills. Secondly, contrary to Byass' InterVA-4 software and the InterVA of R, this metadata driven module allows VA 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. This distribution of causes of death has shown that over 75

Conclusion

As the WHO Verbal Autopsy questionnaires used by several INDEPTH research centres can be adapted to take into account local contextual factors, 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 recommendations. The authors 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. The 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 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.

List of abbreviations

API: Application Program Interface;
 DMBS: Database Management System.
 DSA: Demographical Surveillance Area;
 CAPI: Computer-Assisted Personal Interviewing;
 CSV: Comma Separate Value;
 CRF: Case Report Form;
 HDSS: Health and Demographic Surveillance System;
 INDEPTH: International Network for the Demographic Evaluation of Populations and their Health;
 GPS: Global Positioning System;
 InterVA: Interpretation of Verbal Autopsy;
 ICD: International Classification of Diseases;
 CoD: Causes of Death;
 GIS: Geographical Information System;
 ODK: Open Data Kit;
 SQL: Structured Query Language;
 XML: Extensible Mark-up Language;
 VA: Verbal Autopsy;
 WHO: World Health Organization;
 XHTML: eXtensible HyperText Markup Language;
 RDBMS: Relational Database Management System;
 MSSQL: Microsoft SQL Server;

Additional Files

Additional file 1 - Verbal Autopsy Instrument

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

Additional file 2 - Details of Microsoft .NET function of VA interpretation created from InterVA of R package

The function of Microsoft .NET of VA interpretation based on InterVA of R package has the following structure:

```
.engine.Evaluate(string.Format("VAOUTPUT<-InterVA(VAINPUT, HIV = '0', Malaria = '1', directory =  
dirPath, filename = '2', output = '3', append = FALSE, groupcode = FALSE, replicate = FALSE)",  
hivp, malariap, filename, outputp));
```

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 of the InterVA 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.

Declarations

Ethics approval and consent to participate

An ethics approval was not required since only anonymized individual data was used to evaluate the software tool. These data were provided by Ouagadougou and Dodowa HDSSs who granted us the access. This was in line with respect to the data protection law.

Consent for publication
Not applicable.

Availability and requirements

- Project name: VAMP
- Project home page: <http://sourceforge.net/projects/vap1>
- Operating system(s): Windows, Linux, UNIX Server systems, Mac Os
- Programming language: ASP.NET(C#)
- Other requirements: NET 4.5 or higher
- License: LGPL
- The sample datasets used during tests operations of the current platform are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by INDEPTH Network through the scholarships in Research Database Management Master degree at Wits University.

Author's contributions

KK developed the software, carried out the data analysis, the evaluation and wrote the manuscript. GN and IM supervised the work through suggestions and recommendations. Both authors have read and approved the final manuscript.

Acknowledgements

The authors appreciate the support made by INDEPTH Network for KOMBASSERE Kouliga to complete this Master programme at the University of the Witwatersrand of Johannesburg in South Africa with this great dissertation.

Author details

¹Universite Joseph KI-ZERBO, Institut Supérieur des Sciences de la Population, Burkina Faso, Avenue Charge Degaulle, Ouagadougou, BF. ²African Union Development Agency (AUDA-NEPAD) and University of the Witwatersrand, South Africa, 230 15th Rd, Midrand, Johannesburg, SA. ³Public Health Department, University of the Witwatersrand, Johannesburg, SA.

References

1. A global assessment of civil registration and vital statistics systems: monitoring data quality and progress. *The Lancet* **386**(10001), 1395–1406 (2015)
2. Leita, J., Chandramohan, D., Byass, P., Jakob, R., Bundhamcharoen, K., Choprapawon, C., De Savigny, D., Fottrell, E., França, E., Frøen, F., et al.: Revising the verbal autopsy instrument to facilitate routine cause-of-death monitoring. *Global health action* **6** (2013)
3. World Health Organization (WHO): ICD-10 Revision (Internet). <http://www.who.int/publications/i/icd-10-revision.aspx>
4. Lozano, R., Alan D Lopez, C.A., Mohsen Naghavi, A.D.F., Murray, C.J.: Performance of physician-certified verbal autopsies: multisite validation study using clinical diagnostic gold standards. *Population Health Metrics* **9** (2011)
5. Desai, N., Aleksandrowicz, L., Miasnikof, P., Lu, Y., Leita, J., Byass, P., Tollman, S., Mee, P., Alam, D., Rathi, S.K., 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. *BMC medicine* **12**(1), 20 (2014)
6. Sankoh, O., Byass, P.: The INDEPTH Network: filling vital gaps in global epidemiology. IEA (2012)
7. Organization, W.H., et al.: Civil registration: why counting births and deaths is important. *World Health Organization* (2014)
8. Murray, C.J., Lozano, R., Flaxman, A.D., Serina, P., Phillips, D., Stewart, A., James, S.L., Vahdatpour, A., Atkinson, C., Freeman, M.K., et al.: Using verbal autopsy to measure causes of death: the comparative performance of existing methods. *BMC medicine* **12**(1), 5 (2014)
9. Fottrell, E., Byass, P.: Verbal autopsy: methods in transition. *Epidemiologic reviews*, 003 (2010)
10. Byass, P., Huong, D.L., Van Minh, H.: A probabilistic approach to interpreting verbal autopsies: methodology and preliminary validation in vietnam. *Scandinavian Journal of Public Health* **31**(62 suppl), 32–37 (2003)
11. Zahr, C.A.: Verbal Autopsy Standards: Ascertaining and Attributing Cause of Death. *World Health Organization*, ??? (2007)
12. Flaxman, A.D., Vahdatpour, A., Green, S., James, S.L., Murray, C.J.: Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. *Population health metrics* **9**(1), 29 (2011)
13. Organization, W.H.: WHO | Verbal Autopsy Standards: Ascertaining and Attributing Causes of death. <http://www.who.int/healthinfo/statistics/verbalautopsystandards/en/> Accessed 2018-06-21
14. King, G., Lu, Y., Shibuya, K.: Designing verbal autopsy studies. *Population Health Metrics* **8**(1), 19 (2010)
15. Network, I.: Indepth standardized verbal autopsy questionnaire. Available from: www.indepth-network.org/core_documents/indepthtools.htm (2003)

16. Tamgno, J.K., Faye, R.M., Lishou, C.: Verbal autopsies, mobile data collection for monitoring and warning causes of deaths. In: *Advanced Communication Technology (ICACT)*, 2013 15th International Conference On, pp. 495–501 (2013). IEEE
17. Benzler, J., Herbst, K., MacLeod, B.: A data model for demographic surveillance systems (1998)
18. Byass, P., Chandramohan, D., Clark, S.J., D'Ambrosio, L., Fottrell, E., Graham, W.J., Herbst, A.J., Hodgson, A., Hounton, S., Kahn, K., et al.: Strengthening standardised interpretation of verbal autopsy data: the new interva-4 tool. *Global health action* **5** (2012)
19. Homan, T., Pasquale, A., Kiche, I., Onoka, K., Hiscox, A., Mweresa, C., Mukabana, W.R., Takken, W., Maire, N.: Innovative tools and openhds for health and demographic surveillance on rusinga island, kenya. *BMC research notes* **8**(1), 397 (2015)
20. Weissman, C.D., Bobrowski, S.: The design of the force.com multitenant internet application development platform. In: *SIGMOD Conference*, pp. 889–896 (2009).
<http://cloud.pubs.dbs.uni-leipzig.de/sites/cloud.pubs.dbs.uni-leipzig.de/files/p889-weissman-1.pdf>
21. Guenther, R., Radebaugh, J.: *Understanding metadata*. National Information Standard Organization (NISO) Press, Bethesda, USA (2004)
22. Li, Z.R., McCormick, T.H., Clark, S.J.: *InterVA4: An R package to analyze verbal autopsy data*. R Foundation for Statistical Computing Vienna, Austria (2014)

Figures

Figure 1

Metadata Driven Process Flow This figure 1 shows the metadata driven process schema and overview of the design process.

Figure 2

Generic Metadata Structure This figure 2 shows the mapping 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

Figure 3

Use Case Diagram This figure 3 shows the use cases diagram model that the system needs to perform.

```

graph TD
    QB[QUESTIONNAIRE BANK] --> FG[Forms generation]
    QB --> UT[Utilitiestools]
    FG --> F[Form]
    UT --> VAFTP[VA Form translation processing]
    VAFTP --> ODK[ODK Data dictionary]
    UT --> S[Selection]
    S --> DVAD[Direct verbal autopsy data entry into generated form.]
    DVAD --> Fm[Formite]
    Fm --> DR[Data recording]
    DR --> VADR[VERBAL AUTOPTSY DATA REPOSITORY]
    VADR --> DTS[Data repository synchronisation]
    DTS --> MR[METADATA REPOSITORY]
    HDSS[HDSS ROUTINE DATABASE] --> DTS
    VADR --> DTS
    VADR --> DTE1[Verbal autopsy data extraction]
    DTE1 --> VI[Verbal Autopsy data interpretation process]
    VADR --> DTE2[Verbal autopsy data extraction]
    DTE2 --> CODM[CODs distribution on Map process]
    VI --> CODG[COD distribution on graph]
    VI --> CDO[Cause of death output]
    VI --> ICDF[InterVA data in CSV format]
    CODM --> CODMAP[CoD distribution on a map]
    CODMAP --> GMA[Google Map API V3]
  
```

Form Field	Field Type	Required	Actions
2A120	Text Box	☐ Required?	Delete
Prompt	Name of verbal autopsy interviewer:		

Form Field	Field Type	Required	Actions
2A140	Text Box	☐ Required?	Delete
Prompt	RECORD THE DATE OF INTERVIEW:		

Form Field	Field Type	Required	Actions
2A130	Text Box	☐ Required?	Delete
Prompt	RECORD THE TIME AT START OF INTERVIEW:		

Form Field	Field Type	Required	Actions
2A100	Text Box	☐ Required?	Delete
Prompt	Name of verbal autopsy respondent:		

Figure 5

Questions Metadata Entry Form extract. This figure 5 shows the form for verbal autopsy questions metadata entry form extract where questions codes are entered by end user to generate the verbal autopsy data entry form

InterVA Interpretation results for each death

ID	MALPREV	HIVPREV	PREGSTAT	PREGLIK	PRMAT	INDET	CAUSE1	LIK1	CAUSE2	LIK2	CAUSE3	LIK3
155	I	I	Indet	0			Acute resp infect incl pneumonia	75				
212	I	I	nfp	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	nfp	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	nfp	63	37		Malaria	84				
1327	I	I	Indet	0			Other transport accident	91				

Figure 6

Verbal Autopsy Data Interpretation Result. This figure 6 shows an extract of Verbal Autopsy dData interpretation result except values of variables ID, MALPREV, HIVPREV that are automatically generated by the function can be exported into Excel 2007 or OpenOffice Calc spreadsheet and earlier versions.

Cause-Specific Mortality Fraction from Verbal Autopsy Data

Figure 7

Cause-Specific Mortality Fraction and the probability of the CoD Distribution. Tis figure 7 shows an extract of the cause-specific mortality fraction and the probability of the cause of death distribution in the population.

Figure 8

Population of CoD distribution. This figure 8 shows that the probabilities for Malaria and acute respiration infection in the HDSS area are higher than the rest of the causes of death.

Figure 9

COD Distribution from Communicable Disease in a Formal Area. This figure 9 shows causes of death distribution of communicable disease in one of an Ouagadougou HDSS formal area, automatically obtained through the verbal Autopsy platform after data entry.

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.		

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 10

Use Case Description. This figure 10 describe meanings of each use case.

Figure 11

Relational Database Model for the VA Data Platform This figure 11 shows the relational database model built based on these uses cases for the VA Platform using InterVA-4 model.

Figure 12

HDSS Sites VA Data Integration Process This figure 12 shows the data integration process in which the system checked data integrity and data quality based on the InterVA Matrix data format requirements.