Text Analytics to Predict Time and Cause of Death from Verbal Autopsies

Samuel Odei Danso

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Leeds
School of Computing

September 2015

Declaration

The candidate confirms that the work submitted is his/her own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Some Chapters of this thesis are based on jointly-authored publications. In publications for which I was the lead author, the co-authors acted in an advisory capacity, providing supervision and feedback. In publications for which I was a co-author, details of the work which was directly attributable to me have been stated. All original contributions presented here are my own.

Chapter 1, 2, 5 section 5.2.1

Ansah M, A.; ten Asbroek, A.; Soremekun, S.; Gyan, T.; Weobong, B.; Tawiah-Agyemang, C.; **Danso, S**.; Amenga-Etego,S; Owusu-Agyei, S , Hill, Z. And Kirkwood, B. R. (2014). Evaluating the implementation of community volunteer assessment and referral of sick babies: lessons learned from the Ghana Newhints home visits cluster randomized controlled trial. *Health Policy and Planning.* 29(2):pp114-127

Kirkwood, B.R.; Manu,A; ten Asbroek, G; Soremekun, S; Weobong,B; Gyan, T; **Danso, S**; Amenga-Etego, S; Tawiah-Agyemang, C; Owusu-Agyei,S; HillZ.(2013) Effect of the Newhints home-visits intervention on neonatal mortality rate and care practices in Ghana: a cluster randomised controlled trial. *The Lancet,* 381(9884):pp2184-2192

Hurt ,L; Asbroek, A; Amenga-Etego, S; Zandoh, C; **Danso, S**; Edmond ,K,Hurt, C;Tawiah, C;Hill, Z; Fenty, J; Owusu-Agyei, S;Campbella, O and Kirkwood B. (2013). Effect of vitamin A supplementation on cause-specific mortality in women of reproductive age in Ghana: a secondary analysis from the ObaapaVitA trial. *Bull Health Organisation*,(911):pp19-27

My contributions: the design and implementation of the data managements systems as required by the design of the study.

Chapter 5 and 6: from section 5.2.1

Danso, S; Atwell, ES; Johnson, O; ten Asbroek, A; Soromekun, S; Edmond, K; Hurt, C; Hurt, L; Zandoh, C; Tawiah, C; Fenty, J; Etego, S; Agyei, S; Kirkwood, B. (2013) A Semantically Annotated Verbal Autopsy Corpus for Automatic Analysis of Cause of Death. *ICAME Journal of the International Computer Archive of Modern English.* 37. pp37-70

Danso, S; Atwell, ES; Johnson, O; ten Asbroek, A; Soromekun, S; Edmond, K; Hurt, C; Hurt, L; Zandoh, C; Tawiah, C; Fenty, J; Etego, S; Agyei, S; Kirkwood, B. (2013). Verbal Autopsy Corpus Annotated with Cause of Death. *In proceedings of Corpus Linguistics Conference Proceedings.* Lancaster, UK [Online]http://www.birmingham.ac.uk/documents/college-rtslaw/corpus/conference-archives/2011/Paper-271.pdf

My contributions: principal author, developed methods and software used in transcription of the corpus, carried out analysis of the corpus and produced drafts of the manuscripts.

Other Authors' contributions: collection and organisation of the annotations of the corpus and providing feedback on drafts and general guidance to paper write-up.

Chapter 7

Danso, S; Atwell, ES and Johnson, O. (2013). A Comparative Study of Machine Learning Methods for Verbal Autopsy Text Classification. *International Journal of Computer Science Issues*.10(6). [Online] http://arxiv.org/abs/1402.4380

My contributions: principal author, developed all algorithms, carried out experiments, analysis of results, and produced drafts of the manuscripts.

Other Authors' contributions: provided feedback on drafts and general guidance to paper write-up

Chapter 8 and 9

<u>Danso, S</u>; Atwell, ES Johnson, O. Linguistic and statistically derived features for Cause of Death prediction from Verbal Autopsy text. *Proceedings of the international conference of the German Society for Computational Linguistics and Language Technology 2013*, *Processing and Knowledge in the web, LNCS/LNAI series, Springer*. 8105:pp47-60

My contributions: principal author, developed all algorithms, carried out experiments and analysis of results and produced drafts of the manuscripts.

Other Authors' contributions: provided feedback on drafts and general guidance to paper write-up.

Acknowledgements

"No Man Is an Island", John Donne

It has been a long journey and when writing this part of the thesis it becomes evidently clear that the long journey has at last come to a successful end. This journey would not have come to this end without the help of many people, whom I have referred to as the ‘enablers’. I have an endless list of people that I am highly indebted to and my appreciation to these people is beyond what words can express.

First and foremost, my sincere appreciation goes to my supervisors: Dr Eric Atwell and Mr Owen Johnson. Eric and Owen ensured from the onset of this research project that I was fully provided with the necessary resources and environment to enable me to function optimally. I am forever grateful for the mentorship and out-of-office-hours responses to my emails which motivated me throughout the research. The prompt and critical feedback they provided on drafts pushed me beyond my boundaries to achieve good quality outputs. This has resulted in numerous journal publications and conference presentations; an experience I was hoping to acquire as part of my PhD.

I am grateful to the supporting team, made up of both academic and administrative staff at the School of Computing. Dr Katja Markert, who further introduced me to the world of NLP and also provided critical feedback at various stages, enabled me to succeed. George, Judi and Teresa, thank you for the efficient manner in which you dealt with all my requests.

I am also eternally grateful to Professor Betty Kirkwood of LSHTM, who identified the potential in me, and made sure she relentlessly supported me to realise this potential. This thesis is a testimony to her faith and support. This also goes for the rest of the research team: senior colleagues and friends at London School of Hygiene and Tropical Medicine and Kintampo Health Research Centre, Dr Poorna Gunasekera (Plymouth University),Dr Sam Newton (Kwame Nkrumah University of Science &Technology, Ghana), Dr Majdi Sawalha (University of Jordan), Mr Ebenezer Asaah (Baxter College, UK) , and Dr Augustinus G. ten Asbroek (Academic Medical Centre, Amsterdam).Dr Seth Owusu Agyei (Director, Kintampo Health Research Centre) and Mr Seeba Amenga-Etego(Head, Computer Department, Kintampo Health Research Centre). I am grateful to you for all your words of encouragement throughout this period.

My appreciation also goes to the Institute of Child Health of University College London for nominating me, and the Commonwealth Scholarship Commission for providing the funding support for this research.

Another group of people I cannot afford not to mention for their enormous contribution to this success are my family and parents. I am grateful to my parents for their support. I am also eternally grateful to my wife, Pat, and my two daughters, Karen and Linsey, for giving me the space to work and ensuring that I successfully completed this undertaking. This thesis is a testimony to your sacrifices.

Abstract

This thesis describes the first Text Analytics approach to predicting Causes of Death (CoD) from Verbal Autopsies (VA). VA is an alternative technique recommended by the World Health Organisation for ascertaining CoD in low and middle-income countries (LMIC). CoD information is vitally important in the provision of healthcare. CoD information from VA can be obtained via two main approaches: manual, also referred to as the physician-review and automatic. The automatic-based approach is an active research area due to its efficiency and cost effectiveness over the manual approach. VA contains both closed responses and open narrative text. However, the open narrative text has been ignored by the state-of-art automatic approaches and this remains a challenge and an important research issue. We hypothesise that it is feasible to predict CoD from the narratives of VA. We further contend that an automatic approach that could utilise the information contained in both narrative and closed response text of VA could lead to an improved prediction accuracy of CoD.

This research has been formulated as a Text Classification problem, which employs Corpus and Computational Linguistics, Natural Language Processing and Machine Learning techniques to automatically classify VA documents according to CoD. Firstly, the research uses a VA corpus built from a sample collection of over 11,400 VA documents collected during a 10 year period in Ghana, West Africa. About 80 per cent of these documents have been annotated with CoD by medical experts. Secondly, we design experiments to identify Machine Learning techniques (algorithm, feature representation scheme, and feature reduction strategy) suitable for classifying VA open narratives (VAModel1). Thirdly, we propose novel methods of extracting features to build a model that predicts CoD from VA narratives using the annotated VA corpus as training and testing set. Furthermore, we develop two additional models: only closed responses based (VAModel2); and a hybrid of closed and open narrative based model (VAModel3).

Our VAModel1 performs reasonably better than our baseline model, suggesting the feasibility of predicting the CoD from the VA open narratives. Overall, VAModel3 performance was observed to achieve better performance than VAModel1 but not significantly better than VAModel2. Also, in terms of reliability, VAModel1 obtained a moderate agreement (kappa score = 0.4)

when compared with the gold standard– medical experts (average annotation agreement between medical experts, kappa score= 0.64). Furthermore, an acceptable agreement was obtained for VAModel2 (kappa score =0.71) and VAModel3 (kappa score =0.75), suggesting the reliability if these two models is better than medical experts. Also, a detailed analysis suggested that combining information from narratives and closed responses leads to an increase in performance for some CoD categories whereas information obtained from the closed responses part is enough for other CoD categories.

Our research provides an alternative automatic approach to predicting CoD from VA, which is essential for LMIC. Therefore, further research into various aspects of the modelling process could improve the current performance of automatically predicting CoD from VAs.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

"As long as you live, keep learning how to live" Lucius Annaeus Seneca

1.1 What is Verbal Autopsy?

Not all deaths that occur annually are medically certified with a Cause of Death (CoD). It is estimated that about 67 percent of the 57 million deaths that occur annually are not medically certified due to weak or negligible death registration systems, and more deaths occurring outside the environments of the health system in low income countries (World Health Organization, 2004). Information about CoD is a means to revealing preventable illness; developing health interventions; and researching for treatment of diseases (Kahn et al., 2000). However, in low income countries there is pressure to find cost effective but still accurate CoD information and the Verbal Autopsy technique is frequently employed to do this (World Health Organization, 2004).

The Verbal Autopsy (VA) is a well-established technique employed in a large number of low income countries, where it is generally followed as standard procedure. It involves interviewing individuals (such as relatives or caregivers) who were close to the deceased, and if possible, those who cared for the individual around the time of death, in order to collect information on any medical history, signs and symptoms and other events that may have led to the individual's death. The interviews are captured on a standard questionnaire document that is then sent for analysis by physicians who agree on CoD. It is worth noting that the VA interviews are mostly carried out in the local languages of the countries in which they are employed, then translated and transcribed onto the VA document into the official language for physicians to review. To ensure standardisation in the review process, a classification system based on the World Health Organisation (WHO) International Classification of Diseases (ICD) coding standards is employed in carrying out the exercise.

The paramount use of VA is to serve as an enabling tool to identify patterns of CoDs at the community or population level in the countries where VAs are predominantly employed (World Health Organization, 2012).

1.2 Rationale and Motivation of Research

Accurate information on causes of death is an essential part of evidence-based health policy formulation, planning and evaluation. It is also based on this information that research on treatments can be prioritised. However, this possibility exists only in places where death registration systems are perfect. VA has been developed to ensure countries with a poor death registration system are also able to contribute to the evidence-based practice which has recently attracted attention from the research community. The research is meant to explore how the processes involved in obtaining death information from VA could be standardized and improved in terms of accuracy and efficiency, with less resources considering the problems characterizing low income developed countries (Soleman et al., 2006). The research efforts can be broadly categorized in three themes:

- VA data collection tools (closed responses versus open-narrative sections of the questionnaire);

- Development of expert algorithms to guide physicians in reviewing VA data collected and assigning the underlying CoD;

- Research into the development of automatic algorithms to predict CoD from VA (Murray et al., 2007).

Presently, there is mixed evidence regarding the best data collection approach to use. The practice thus far has been the use of a closed response and open narrative text to elicit VA information for analysis and CoD determination. This agreement was reached by the WHO in consultation with experts and researchers in VA (Soleman et al., 2006). Furthermore, several studies conducted by various researchers point to the fact that information derived from the open narrative text has contributed to the overall CoD prediction by physician review when validated against a computer algorithm (Freeman et al., 2005). This open narrative text information is not taken into account by the current automatic approaches (Byass et al., 2010, King et al., 2010), and this has resulted in their criticisms and non-acceptance (Gajalakshmi and Peto, 2006).

Research into automatic approaches to predicting CoD is well advanced. Validation studies have demonstrated the feasibility of adapting automatic approaches to CoD prediction with promising results when compared with the manual approaches such as the physician review (Byass et al., 2010, King et al., 2010). However, the level of accuracy of these automatic approaches needs to be improved in order to be able to compete favourably against the manual method.

Additionally, there is currently a knowledge deficit about the potential usefulness of the information contained in the open narrative part of VA in computational modelling. The reason for this shortfall is partly due to the difficulty involved in obtaining the open narrative text in a machine readable format and also the lack of methods and algorithms developed for extracting the information embedded in the open narrative text, which this research attempts to address. The knowledge to be obtained in this research will potentially enrich the ongoing discussion regarding the justification of collecting VA information in both closed responses and open narrative text formats (Soleman et al., 2006).

1.3 Research Aims and Objectives

The aims and objectives of this research can be summarised as follows:

1.3.1 Aims

- To determine the feasibility of automatically predicting CoD from VA open narrative text.

- To determine the extent to which information derived from the narrative improves on the accuracy of prediction over methods based on closed responses only.

1.3.2 Objectives

- To build a corpus of VA open narrative text to serve as a resource for research.

- To investigate Machine Learning methods and techniques suitable for the VA domain.

- To investigate the features and methods of extraction suitable for predicting CoD from VA open narrative text.

- To disseminate these findings through journal publications and conference proceedings.

1.4 Novelty, Originality and Contributions

1.4.1 Resource and Tools

- This research has built the first ever corpus of VA open narrative text to be used as a resource for language research.

- A software tool which was developed and used in transcribing the VA corpus has been made available to VA researchers and it is currently being used by the London School of Hygiene and Tropical Medicine as well as The Centre for Global Health Research at the University of Toronto, Canada.

1.4.2 Methods

- This research has established suitable methods of collecting and transcribing the VA open narrative text.

- The research has developed novel methods for predicting CoD from VA open narrative text.

- This research carried out a comparative study of Machine Learning approaches and has identified a suitable approach to analysing and predicting CoD from VAs.

1.4.3 Impact

The impact of this research is observed in two domains:

Natural Language Processing (NLP) and Machine Learning Research Community

This research applied a combination of Corpus and Computational Linguistics, Natural Language Processing (NLP) and Machine Learning approaches to automatic analysis and predicting CoD from VA open narrative text. This is the first research to employ that approach in this area. The feasibility of this approach has been established with promising results, and formed the basis for further NLP and Machine Learning research within the context of VAs. For example, in Chapter Seven we discuss well-established NLP methods employed in extracting novel lexical and phrasal feature patterns from VA text for classification. Furthermore, in Chapter Five we evaluate the performance of current state-of-the-art Part-of-Speech (PoS) taggers on the VA dataset, which suggests additional work required to improve on their performance. This is a potential area for NLP researchers with particular focus on PoS taggers.

Verbal Autopsy Research Community

Although it is established within the VA research community that the open narratives of VA are useful to medical experts during the manual process of analysing and determining CoD, little is known about the feasibility of employing an automatic approach to predicting CoD from these narratives of VA. This gap in knowledge has resulted in limited use of the narratives despite the fact that both closed responses and open narratives text are collected. We demonstrate in Chapter Seven the feasibility of automatically predicting from the open narratives. Furthermore in Chapter Nine we evaluate the usefulness of information from both closed and narratives from the perspective of automatic approaches and what these findings mean to VA researchers and practitioners.

1.5 Scope of the Research

Although the corpus built for the research comprises of data collected from both adult women and infant death, the experiments and the results presented are based on the infant corpus. However the generalisability of our methods in theory should be applied to the adult women sub-corpus. Future work will explore this in detail once the corpus has been made publicly available, something which is currently in process.

1.6 Thesis Outline

The rest of the thesis is organised as follows:

Chapter Two is comprised of a survey of the VA field. A brief overview of the field is given, which describes the processes involved in gathering VA information; the places of use; and the various levels of users of this information. The Chapter also presents a literature review conducted on the various automatic approaches that currently exist in carrying out analysis and predicting the CoD from VA, which informs the formulation of our research. Finally, Chapter Two concludes by giving an overview of our approach to analysis and how this differs from the existing approaches to analysis and predicting causes of death from VA data.

Chapter Three is concerned with the approaches to Text Classification. The processes and methods employed in carrying out the Text Classification tasks are explored. We particularly focus on the feature engineering process and the various automatic techniques available for extracting and representing features for classification of textual data. The Chapter concludes with a survey of the applications of Text Classification methods within the biomedical domain since this research is situated within the biomedical domain.

Chapter Four explores the various Machine Learning algorithms employed in performing the classification tasks. Furthermore, the Chapter discusses the evaluation metrics that are employed in determining the performance of a given Machine Learning method. We then identify the overlaps between the evaluation metrics employed in the Machine Learning and biomedical domains. The Chapter concludes by looking into some of the issues that tend to affect performance of Machine Learning algorithms in the context of Text Classification.

In Chapter Five, the focus is on data collection and our VA corpus. This Chapter starts by describing the surveillance system in Ghana which provided the source for the VA data and the Data Management System employed to deal with the inherent complexities involved in managing data emanating from a relatively large population. It then proceeds to describe the methods employed in building the open narrative text corpus for this research. The Chapter further describes principles and strategies employed in managing the complexities associated with the cause of death classification schemes employed in this research.

In Chapter Six we conduct and consider a Corpus Linguistics study of our VA corpus by analysing the content, discussing the language issues and the computational challenges associated with the corpus in the context of Machine Learning algorithms for Text Classification.

Chapter Seven reports on the initial experiments carried out to identify the appropriate automatic techniques suitable for the problem under investigation. In this Chapter we propose an approach called ‘locally-semi-automatic' as a feature reduction strategy to identify keywords, referred to in this thesis as Discriminative Word Units (DWU) for Text Classification of a noisy text. Finally a recommendation is made based on the experiments, which is subsequently employed in the methods that are discussed in the next Chapter.

Chapter Eight is concerned with the engineering of the feature space of VA open narrative text. It describes the methods and experiments carried out to extract features within the VA textual documents in order to achieve optimum classification accuracy. The Chapter concludes by conducting a comparative analysis to identify the novel features for predicting causes of death from VAs.

Chapter Nine extends the discussions to include findings from experiments that considered the closed response part of the Verbal Autopsies and other domain knowledge information that were not taken into account in the experiments described in Chapter Seven and Eight.

Chapter Ten provides a summary of the work carried out in this research. It highlights the major findings, those being an alternative approach to the existing automatic approaches for finding CoD from VA. The limitations of this research are also discussed. The Chapter concludes by outlining a road map of future work that could be explored in order to enhance the performance accuracy achieved by the prediction model.

Chapter 2

Research Background

"Necessity is the mother of invention", Plato

2.1 Introduction

Considering the importance of cause of death information discussed in Chapter One, the drive to develop innovative ways to gather this information is strong. In this Chapter we look into the historical perspective of VA as one solution for capturing cause of death information and how it has evolved to date. The methods of analysis employed, and how the cause of death information is obtained from VA will also be explored. Finally, we describe our proposed approach for analysing and predicting cause of death from VA

2.2 Historical Perspective on Verbal Autopsy

The idea of VA has been around since before the beginning of the 19th century, when rigorous death registration systems were not in existence in most European countries. For example, in the 17th century English doctors employed the principles of VA: persons were designated to visit homes of deceased people to elicit information about the nature of their deaths (Garenne and Fauveau, 2006). More recently, European countries such as Sweden and Ukraine have employed VA as a means of ascertaining causes of death. In Sweden, cases of maternal deaths among migrants to Sweden were identified through VAs (Elebro et al. 2007). In Ukraine, VA was employed to understand and explore the circumstance that led to the death of individuals with diabetes (Telishevka et al. 2001). Most developed countries employ reliable methods of determining causes of death. Countries in Africa and Asia started to adopt the VA approach during the 1950s and 60s due to a lack of the comprehensive health service structures needed to support more modern methods. A systematic approach, instituted by physicians in India who conducted interviews to determine the cause of death, became known by the term *Verbal Autopsy* and this has now become the term most widely used for this method (Garenne and Fauveau, 2006).

In 1956 the concept of VA as information captured through "lay reporting" of health data was introduced through a publication by Dr Yves Biraud (World Health Organisation, 2012). This encouraged the World Health Organisation

(WHO) to formally support the use of lay reporting of health information by non-medical professionals in the 1970s. The idea was developed further and led to the establishment of a standard form known as the "lay reporting form" in 1975. The late 70s and early 80s saw the introduction of wide scale use of Verbal Autopsies in the form of different questionnaires and their use in different settings and surveys (World Health Organisation, 2012).

The coverage and research interest has deepened in recent years and resulted in the active research and development of various aspects of the VA process. Since 2007, the WHO has played a leadership role and worked with VA researchers and practitioners to standardise the VA data collection tools and reporting.

2.3 The Verbal Autopsy Questionnaire.

Figure 2.1 and 2.2 show samples of a section of a VA questionnaire, with a completely filled sample found in Appendix B. Chapter One established that VA data included closed responses and open narrative, and Figure 2.1 shows the closed questions that are carefully designed, based on WHO knowledge of expected symptoms during pregnancy. The value *1* represents y*es; 2* represents *no;* and *8* represents *Not Known (NK)* response. Figure 2.2 also shows an example of an open narrative text response to an interview describing events during the pregnancy of a mother who lost the child, as captured by the interviewer.

6.1. Pregnancy

6.1.1. Maame no wo nkwadaa dodoô sân, sâ wokan wôn a ôwo twene ka ho a, ansa na ôrewo akwadaa yi?

1. None	2. One	3. Two	4. Three	5. Four	6. Five or more	8. NK	PARITY

Question							Code
6.1.2 Na maame no awô nteteâ paneâ no nyinaa bi ansa na ôreduru ne mpanin fie so a na n'yinsân wei ka ho?......	1. Yes	2. No	8. NK				VACCINATEFULL
6.1.3. Nteteâ oaneâ no dodoô sen na w'awô?......	1. One	2. Two	3. Three	4. Four	5. Five or more	8. NK	VACCINDOSE
6.1.4. Maame no wôô asesene paneâ no bi wô ayinsân wei mu?......	1. Yes	2. No	8. NK				TTOXOID
6.1.5. Asensene paneɛ dodoɔ sɛn na wowɔɔ no saa anyinsɛn no mu? [00 = NONE, 88 = NK, ASK TO SEE ANY MEDICAL RECORDS, YELLOW CARD]......	0	0					TETTOXD
6.1.6. Efirii sɛ yɛwoo wo, wowɔɔ asensene paneɛ sɛn ansa na worenyinsɛn wanyinsɛn no? [00 = NONE, 88 = NK, ASK TO SEE ANY MEDICAL RECORDS, YELLOW CARD]......	8	8					TETTOXB
6.1.7. Mprɛ dodoɔ sɛn na ôde n'anyinsɛn no kɔɔ ayaresabea maa wɔhwɛɛ anyinsɛn no? [00 = NONE, 88=NK, ASK TO SEE ANTENATAL CARE RECORD, EXCLUDE ILLNESS]	0	0					ANC
6.1.8. Saa ayinsân wei mu no, wo nyaa mogya borosoô?	1. Yes	2. No	8. NK				HIBPPREG
6.1.8.1. Wei sii bâyâ bosome miensa a âtwatoô ansa na wore wo?......	1. Yes	2. No	8. NK	9. NA			LATEPREGBP
6.1.9. Abere a na woyim no, mogya tuu wo bâberec firi w'ase?......	1. Yes	2. No	8. NK				VAGBLEED
6.1.9.1.Wei sii wô bosome nsia edi kan wô ayinsân no mu anaa?	1. Yes	2. No	8. NK	9. NA			EARLYVAGBLEED
6.1.9.2.Wei sii wô bosome mneinsa âtwatoô wô ayinsân no mu anaa?......	1. Yes	2. No	8. NK	9.NA			BLEED3MTHS
6.1.10. Wô bere a na wo yim no mu no, nsuo bi guu wo firi w'ase pân a na panpan wô mu?	1. Yes	2. No	8. NK				VAGDISC
6.1.10.1.Wei sii wô bosome miensa âtwatoô wô ayinsân no mu anaa? (Last 3 months)......	1. Yes	2. No	8. NK	9. NA			DISCHARGE
6.1.11. Wô w'ayinsân mu no, âdu bere bia na w'aniso yâ wo wisiwisi?......	1. Yes	2. No	8. NK				BLURVIS

Figure 2.1 closed questions: pregnancy section of VA questionnaire

FIRST ASK "Wobetumi aka biribi afa berɛ a na wonyem akwadaa no ho akyerɛ me?"

[redacted] never attended Antenatal care clinic (ANC) during her pregnancy.
She dose not felt sick during the pregnancy.
~~She also attended Antenatal care clinic (ANC)~~
for. She [redacted] delivered prematurely at seven (7) months.

Figure 2.2 open narrative text response: pregnancy section of VA questionnaire as captured by interviewer.

2.4 Levels of Users of Verbal Autopsy information

Byass et al (2007) classifies 5 groups of users of the information derived from VA data.

2.4.1 National and International Bodies

The information required from VA by this group of users is the global and national cause-specific mortality estimates. This information should be in a standardised format, which is usually obtained using the ICD coding system. This allows standard comparability of mortality estimates over time from different sources and places, using complex models.

2.4.2 Local Public Health Managers

This group of users needs information on the rankings of causes of death in order to monitor trends over time and evaluate any public health interventions. Relatively simple models are employed to obtain the required information.

2.4.3 Epidemiologists and Health Services Researchers

This group of users needs VA information to analyse mortality patterns in order to understand a given context. This is usually done with reference to specific populations, which is achieved through consistent approaches to analysing VAs to determine cause of death.

2.4.4 Institutional Managers and Clinical Auditors

This group takes interest in monitoring trends of cause of death within institutions and healthcare systems for management purposes. This level of cause of death is usually obtained through physician certification and medical record reviews.

2.4.5 Medical and Legal Practitioners

Cause of death information is required for this group of users to carry out post event consequence assessment of the death, which is usually done at the individual level.

The above group of users have further demonstrated the importance of VA CoD information. It is therefore imperative to explore methods that can accurately and efficient generate this information to assist these users in carrying out their responsibilities as required.

2.5 The Verbal Autopsy Process

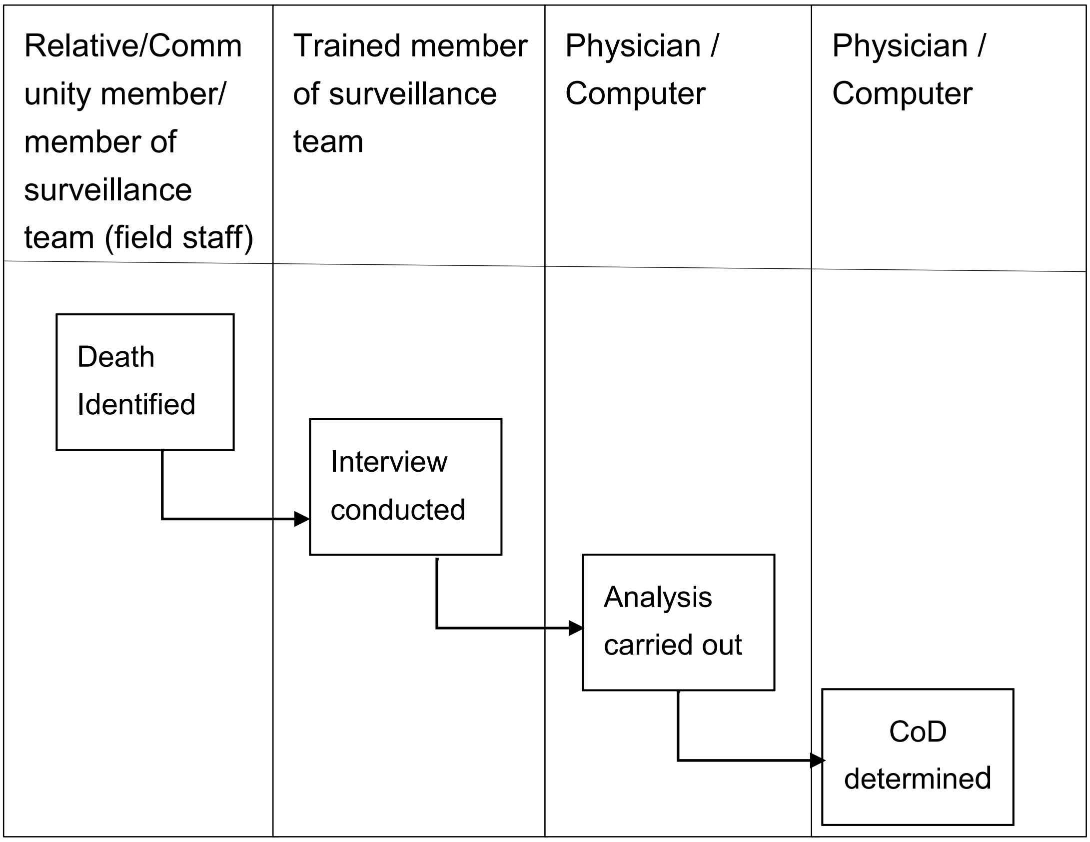


Figure 2.3 swim lane diagram showing the Verbal Autopsy process

Figure 2.3 shows a swim lane diagram that demonstrates the VA process. It begins when a death is reported in the community. This is done mostly through a relative of the deceased, a community member who heard about the death in the community or a member of the surveillance team or field staff who normally resides in the community. Once the death is discovered through some form of surveillance system or a relative reporting the death to an institution of interest, such at the local health authority or a research institution interested in causes of death, an interview is conducted. This would be done by someone trained by that institution to record the history of events that led to the death. The interviewee is typically a relative who was close to the deceased. The interview is conducted using a locally adapted version of the WHO designed VA questionnaire. The analysis phase occurs when the information obtained from the questionnaire is analysed using various approaches (see next section) and then CoD is determined. Aggregated statistical analysis can then be carried out to establish mortality trends such as *Cause Specific Mortality Fractions* (CSMF) in the population.

2.6 Approaches to Verbal Autopsy Analysis.

Approaches to VA analysis and CoD determination from VA can be classified into three main categories (King and Lu, 2008). The first approach is the physician review. This is where physicians manually review the VA data and assign the CoD. The second approach is the use of an expert algorithm, which involves using a consensus of physicians' judgments to construct algorithms such as a decision tree to serve as a guide for determining CoD, and the third involves the use of automated methods to determine the CoD (Byass et al., 2006).

Typically, information gathered using VAs are captured on paper using standard questionnaires which are then passed to physicians who review them to determine the most likely cause of death (Byass et al., 2006). This approach is referred to as a Physician Certified VA (PCVA). The standard practice of PCVA worldwide has been the use of a minimum of two physicians to give two or more independent assessments of each VA, even though there is some evidence to suggest that one physician may be enough for this process (Joshi et al., 2009). The PCVA approach is characterised by several limitations: high cost; inter-physician reliability; repeatability; and time consumption (Byass et al., 2010). The cost of manual review and assignment of cause of death to VA documents has not been formally evaluated. However, this exercise may be equated to assigning International Classification of Diseases (ICD) codes to clinical documents, the manual coding of which had an estimated cost of approximately $25 billion per year in the United States alone (Lang, 2007). The problem is compounded where there are shortages of medical personnel, as is generally the case in places where VAs are used.

Consequently, there has been a growing interest in research surrounding the use of automatic approaches to classify causes of death (King et al., 2010, Byass et al., 2010). Having conducted a systematic search of the literature of databases such as PubMed and Google scholar with the keywords Verbal Autopsy and manual review of abstracts to initially determine the relevance after a thorough review was carried out, the next section briefly discusses the current state-of-the-art automatic approaches to VA analysis.

2.7 The State-of-the-art Automatic Approaches

Automatic approaches to the determination of CoD have been an active area of research since Quigley et al. (1996) explored automatic approaches to CoD determination in VA. Their study involved the use of logistic regression to predict CoD. It was a validation study that aimed to compare the performance in terms of accuracy between the expert systems approach and their statistical approach. They used data from 295 children who had died in health facilities. The relatives of these children were identified and interviewed using a VA questionnaire based on a predefined set of closed questions. The CoD of these children were known from the health facility. The symptoms data, which has categorical responses for *Yes* and *No*, and the confirmed CoD were used to train a logistic regression algorithm that could then be used to predict CoD in the community. The basic assumptions made in the study were that the symptoms collected within the community should have the sensitivity and the specificity required to determine the CoD. Also, both hospital and community had the same CoD predictive features and therefore could be treated as one. Although their results showed a promising 71% accuracy for selected diseases, their assumptions may not hold practically as argued by King and Lu (2008). Hospital data may not be available in most settings where VAs are employed. Furthermore, the VAs were dependent on the ability of the respondent to recall symptoms which may not have had the same sensitivity and specificity required by the method compared to hospital data obtained through clinical diagnosis or examination.

King and Lu, (2008) proposed another automatic method, which was built on the previously discussed approach(Quigley et al., 1996) by dropping the assumptions discussed above. King and Lu argued that the possibility existed to use hospital CoD data to predict CoD in the community, provided the systems in both populations were the same. They validated their model using data collected from China and Tanzania. The Tanzania dataset had 1261 records from the hospital and from the community, 51 symptoms of closed questions of categorical responses of *Yes* or *No* and 13 CoDs. Even though both datasets had medically certified CoDs, the community CoDs were removed from the dataset and were used to validate the results of their model.

The China dataset had 2822 records from the hospital where CoDs were known. The hospital data was used to train the model to predict the CoD from the community data. Although they claimed to have better performance in estimating the CoDs this was not clearly reported in terms of the performance accuracy achieved. They however argued that their approach was good in estimating CSMF at the population level. Again, the assumption of this approach would be difficult to meet in practical terms; the assumption was based on the availability of accurate death information from the health facility, which is not readily available in the setting where VA is mostly employed.

Recent developments have seen a shift from the use of hospital data in developing VA prediction models to the use of community data for both modelling and evaluation of VA models (Fottrell et al., 2007; Byass et al., 2006). The approach was recently validated against the physician review and it is reported to have comparable performance (Byass et al., 2010). The validation study involved VA data collected in South Africa containing 6153 cases. The CoD of these had already been determined using physician review. The probabilistic model is based on Bayes' theorem as expressed in the equation below, and had a software implementation called interVA (Fottrell et al., 2007, Byass et al., 2006).

$$P(c|\, i) = \frac{P(i \mid c) \times P(c)}{P(i \mid c) \times P(c) + P(i \mid !c) \times P(!c)}$$

where: c = cause of death; *i* = presence of a given symptom; P(!*c*) = probability of not *c*

Through this equation the probability of each symptom ($i_1 \ldots i_n$) and each possible cause of death ($c_1 \ldots c_m$) for all deaths for a given population can be determined. This means, for any given case, the probability of c_{m-1}, given the symptom i, is initially the value found among all deaths in a given population (prior-probability), which is estimated by human experts. For example, if *c* is death from *Malaria*, and *i* is *Fever*, then the probability of dying from Malaria given Fever can be related to the probability of those who actually die from Malaria whilst either having or not having Fever. The probabilities of the symptoms are computed in order to predict the probable CoD. Only the symptoms with responses *yes* are considered in this model. The results obtained from the study demonstrated good performance. The model was reported to have achieved 83.3% as opposed to 88.2% achieved by the physician approach for the top-10 diseases. The problem however with this approach is the over reliance on medical experts for the estimates of the prior-probabilities(Fottrell et al., 2007, Byass et al., 2006). The medical experts may rely on their knowledge of the prevalence of diseases and symptoms and other assumptions that may not hold in the context of the underlying data, which could cause unrealistic and inaccurate predictions of the model. For example, in malaria endemic settings, malaria symptoms that tend to overlap with other causes may be overestimated for Malaria. An ideal approach would be to have a data driven based method that does not rely on estimates introduced outside the data. The approach being proposed in this thesis differs from the interVA approach because it is data driven.

Another probabilistic approach proposed by Murray et al.(2007) was a hybrid that combined the two methods discussed above. This approach is known as the Symptom Pattern method. It uses data from both the hospital and VA. From the hospital data, the properties of each symptom are determined by computing the probability of responding 'yes' to each symptom given the true CoD. Unlike the previously described method by Byass et al (2006), which uses CoD obtained from VA, this method tended to depend on CoD obtained from hospitals to estimate the probabilities of symptoms contained in the hospital data, which would then be applied to VA data obtained from the same population of interest in order to satisfy the assumption by King and Lu (2008) discussed above. They reported that using this as an input to estimate the CoD at the individual level looked promising using a dataset obtained from China. Their model correctly predicted 83% of the cases compared to 66% achieved by physician review. However, this approach is heavily dependent on the availability of hospital data to estimate rates of symptoms values in populations to serve as input into the model, which again is impractical in most settings where VA is practiced.

Another automatic approach is the clustering-based method of prediction of CoD. Bailly-Bechet et al.,(2009) in their experiment applied a hierarchical clustering algorithm to a 2039 record VA dataset, where the individual CoD were known. The dataset had 47 either *yes*, *no*, or *do not know* responses to symptoms from closed questions. However, data with *do not know* responses were eliminated from the data set before clustering. The result of this was not reported but the authors argued that this framework could be useful for *active learning* where further investigation is needed to determine CoD.

James et al. (2011) have recently proposed another automatic method known as the Tariff method as mentioned earlier. The Tariff method strives to identify signs and symptoms in the VA data that are significantly indicative of a given cause of death. This is based on a score which is assigned to each symptom based on the response pattern observed in the VA data. These scores are then added up for each CoD. The CoD which obtains the highest score is predicted as the CoD for that case. This is then validated against the PHMRC dataset, which contains about 12535 VAs collected from four countries (Murray et al., 2011). The Tariff method is reported to have achieved an accuracy of 44.5% in adults, 39% in children, and 23.9% in neonatal infants. The advantage of this approach as argued by James at al. (2011) is the transparent nature of the process of assigning scores in the decision making process of determining the CoD. Unlike the other automatic approaches, which are considered "black box", this method is considered to be transparent and allows physicians to observe the scores assigned to symptoms of a given cause of death. However, the process of assigning scores to the symptoms by physicians could still be compromised due to the manual process involved, which could also lead to inaccuracy in estimation as a result of overestimating or misunderstanding a given symptom. For example, high Malaria endemic settings could result in overestimating the scores being assigned to Malaria related symptoms and consequently inaccurate prediction.

Flaxman et al. (2011) have also proposed Random Forest (Breiman, 2001) as another automatic approach to predicting CoD from VA. The method is based on a Decision Tree which is generated from the root node of the tree by random resampling of the training dataset. At each node, the algorithm selects a random subset of signs and symptoms to consider branching on, and then branches on the one that best distinguishes between the causes of death relevant to that node. This process creates a pair of causes of death, and generates 100 decision trees of pairs of causes of death. Subsequently a “pairwise coupling”, a voted, ranked and normalised algorithm is employed to determine the probable cause of death. This method is also based on the closed response part of VA and was validated against the PHMRC dataset, which contains about 12535 VAs collected from four countries (Murray et al., 2011). They argue that their approach performed competitively when validated against the physician review.

Flaxman et al. (2013), have recently proposed another automatic method based on the Ensemble concept in Machine Learning, which seeks to combine some of the previously discussed methods: the *Simplified Symptom Pattern* (Murray et al., 2007); *Random Forest* methods, developed by Flaxman et al. (2011); and *Tariff* methods. Each of these methods is made to assign a possible CoD after which a voting process is carried out to determine the majority vote, which is then considered to be the "true" CoD for a given case. A validation study of the *Ensemble* method was carried out using data created by the Population Health Metrics Research Consortium (PHMRC). The results suggested this approach tended to be superior to the individual methods, with individual disagreement between methods at 16% and overall unanimous agreement among the three methods at approximately 47%.

Table 2.1 provides a summary of the current state-of-the-art automatic approaches described above with data sources used. As the table demonstrates a variety of methods have been explored using a variety of data sources as described above. A typical VA contains both closed responses and open narrative information. However, as was established from the survey on the automatic approaches published so far discussed above, all the approaches focused on only the closed response part of the VA. The information contained in the open narrative is ignored by the automatic approaches. Meanwhile, the traditional approaches such as the physician review and expert algorithms have access and make use of both the closed responses and information contained in the open narrative (Soleman et al., 2005). It is therefore imperative to explore automatic methods that can make use of all available information in order to have a better basis for validation and performance comparisons between traditional and automatic approaches.

Table 2.1 Summary of state-of-art automatic approaches to Verbal Autopsy analysis

Authors	Method	Input data
Quigley et al. (1996)	Logistic regression	Closed response part based on hospital data and tested on WHO VA based questionnaire.
King and Lu, (2008)	Probabilistic	Close response part based on simulated hospital data and test on data collected using WHO questionnaire.
(Fottrell et al., 2007; Byass et al., 2006)	Probabilistic (InterVA)	Closed response part based on WHO VA questionnaire only
Murray et al.(2007)	Probabilistic (Symptom Pattern)	Closed response part of a combination of hospital and VA data PHMRC dataset and questionnaire
Bailly-Bechet et al.(2009)	Clustering	Closed response part with data using WHO VA questionnaire
James et al. (2011)	Symptom scoring and ranking (Tariff)	Closed response part of VA based on PHMRC dataset and questionnaire
Flaxman et al. (2011)	Random Forest	closed response part of VA based on PHMRC dataset and questionnaire
Flaxman et al. (2013)	Ensemble - combined Symptom Pattern (Murray et al.(2007), Tariff (James et al. (2011) and Random Forest (Flaxman et al. (2011)	Closed response part of VA based on PHMRC dataset and questionnaire

2.8 Our Approach

Our research is motivated by the belief that an automatic approach that is able to take advantage of information contained in both closed responses and the open narrative should result in a robust and relatively better performing approach. The research has been formulated as a Text Classification problem and we seek to classify the VA according to CoD categories. We have employed Corpus and Computational Linguistics (CL), Natural Language Processing and Machine Learning approaches to identify various features which can be used to classify the VA documents. These approaches are deeply rooted in the sub-disciplines of Artificial Intelligence and are briefly described below.

2.8.1 Corpus and Computational Linguistics

Corpus Linguistics as described by Wallis and Nelson (2001), is the study of language as expressed in samples of the language. This is achieved through the analysis of the samples with the main aim of obtaining linguistic knowledge. For example Corpus Linguistics allows grammatical analysis of words in a text, classifying these words by assigning grammatical labels (verb, noun) to them, referred to as Part of Speech Tagging. These can then be used to construct dictionaries and thesauri by lexicographers (Church and Hanks, 1990).

Computational Linguistics is considered to be a sub-field of Artificial Intelligence, and tends to extend the work of Corpus Linguistics via the application of the understanding of linguistics phenomena and knowledge gained through Corpus Linguistics to develop automatic models. These models can use either a statistical or rule-based approach, which may involve specialised knowledge from other disciplines such as Linguistics, Mathematics and Computer Science.

2.8.2 Natural Language Processing

Natural Language Processing (NLP) is also considered an Artificial Intelligence approach which is mainly concerned with the creation of systems that allow computers to process natural language in order to facilitate human-computer interactions. NLP as a discipline has been around since the 1950s and its evolution has been discussed by Bates (1995). Bates demonstrated the shift in paradigm from an intuition based approach to an increased emphasis on a corpus-based one, which creates an avenue for the combination of NLP with corpus-based methods to language modelling.

2.8.3 Machine Learning

Samuel (2000) defines Machine Learning as an Artificial Intelligence discipline which aims to equip computers with the ability to learn without being programmed. In other words, a computer is considered to have learning capability if the computer is able to learn from experience (*E*), which is normally, based on previous data, obtained from the domain of interest (training data) with respect to some task (*T*), and some performance measure (*P*), if its P on *T* improves with *E* captured from the training set. Machine Learning is driven by learning algorithms which are mathematical functions with a set of parameters (Pereira et al., 2009).

There are numerous factors that determine the choice of Machine Learning algorithms, and this includes the task being performed. The task is categorized in Machine Learning terms as: *Supervised Learning* (Witten and Frank, 2005) , *Unsupervised Learning* (Kotsiantis et al., 2007), and *Semi-supervised Learning* (Bilenko et al., 2004). The method of learning being employed in this research is supervised learning, which allows the algorithm to learn from examples provided by human experts in the form of training examples. Supervised learning algorithms would be capable of learning from the VA data obtained through application of Corpus and Computational Linguistics and NLP techniques in order to predict CoD from VA. The next Chapter discusses classification-based learning applied to biomedical text.

2.9 Summary

This Chapter started by providing a brief historical perspective on VAs. Various groups of users of VA information were identified and the kind of VA information required by those groups of users was also discussed. A brief overview of the entire VA process was also given, which consequently demonstrated the context in which this research is situated in the overall process.

The Chapter also conducted a survey of the automatic approaches to determining CoD from VA published in the literature to date. The survey identified issues related to each of the various approaches. Models that rely on hospital data for model development were identified to be problematic since the settings where VAs are employed may not have hospital data readily available. Also, over reliance on the hospital data for signs and symptoms as proxy for modelling demanded accurate signs and symptoms information from the VA which could be problematic to obtain in VA settings since VA is based on the memory recall of the respondent and not clinical or diagnostic information. Thus, the most pragmatic approach for developing automated methods for VA was the use of VA data which some of the existing methods have begun to explore. However, these methods' tendencies to rely on manual approaches to determine the prior probabilities of the models, as observed from the survey, can also be problematic. For example, inaccurate estimates could lead to inaccurate predictions. Also, as observed from the survey the existing approaches have not explored the open narratives part of the VA data, and this served as a motivation for our research.

An overview of the methods being proposed in this thesis has been described. A clear distinction was drawn and a justification was given to demonstrate how the approach being proposed in this thesis differs significantly from existing approaches.

Chapter 3

Classification of Text: The Biomedical Domain

"Everything you can imagine is real", Pablo Picasso

3.1 Introduction

Chapter Two briefly summarised the rationale for using supervised learning as the approach to be employed in this research. Machine Learning algorithms have the capability to learn from data prepared by human experts. This approach to learning falls into two categories: regression and classification. The difference between these two supervised learning approaches is based on how the problem is formulated. The regression approach is concerned with a learning algorithm predicting a numeric continuous value as an output. The classification-based learning on the other hand is concerned with a learning algorithm predicting a categorical output, which suits the task of predicting causes of death from VA being explored in this thesis. This research is therefore formulated as a Text Classification problem.

Text Classification tends to depend on information obtained from the documents being classified. In this Chapter we explore the process of Text Classification and the techniques that are employed in the classification process. We then further survey the biomedical domain to identify related research works that have successfully applied these techniques.

3.2 Text Classification Process Overview

Text Classification (TC) by definition is an automated process of assigning textual documents to a set of predefined categories (Sebastiani, 2002). This process has seen unprecedented growth in interest and research due to the abundance of documents available in textual format. The process is cross-disciplinary in nature, as it encompasses several subfields under the umbrella of Computer Science: Natural Language Processing (NLP), Machine Learning, Pattern Recognition, and statistical theories (Mitchell, 1997), which are augmented by Corpus and Computational Linguistics as explained in Chapter two.

Formally, TC could be described as a task of assigning a categorical value C_v to each pair of $\{d_j, c_i\}$ € D x C, where D represents the textual documents from a given domain D = $\{d_j,\ldots,d_z\}$ and C= $\{c_1,\ldots,c_m\}$ represents the set of predefined categories that a given document d_j could belong to. The value C_v is determined by a function, $F_{(x)}$ for $\{d_j, c_i\}$, to suggest that the document d_j should be assigned either c_1 category, otherwise d_j should be assigned another member of the set of categories C. $F_{(x)}$ could either be determined by hand crafted rules or a Machine Learning algorithm. Additionally, to classify a document into a given category c_i, given D with a set of unique features F of D, each document d_i € D, has to be represented as a vector $Vd_i = \{v_1,v_2,\ldots v_{IFI}\}$, where |F| is the length of the vector with its s^{-th} dimension Vd_{is} containing a quantified information q_s, q € F that best describes d_i (Jurafsky et al., 2000). The quantification of the feature information is based on representation schemes, which are discussed later in this Chapter. The features could be derived from the content of the d_i obtained through tokenisation by dividing text of the d_i into contiguous characters words, phrases, or symbols that are contextually meaningful based on the problem or targeted application.

There has been a continued effort by the research community towards the aim of improving the performance accuracy of Machine Learning classification algorithms by exploring various subfields. This is due to the fact that numerous factors tend to determine the performance of a given algorithm, and these include: the data and domain; Machine Learning algorithm; and the features and their representation schemes employed in the process of building a classifier for a classification task (Jurafsky and Martin, 2000).The next Chapter will discuss some of the techniques employed in developing a classifier. However, as shown in Figure 3.1, prior to the application of a given algorithm there is the feature engineering phase.

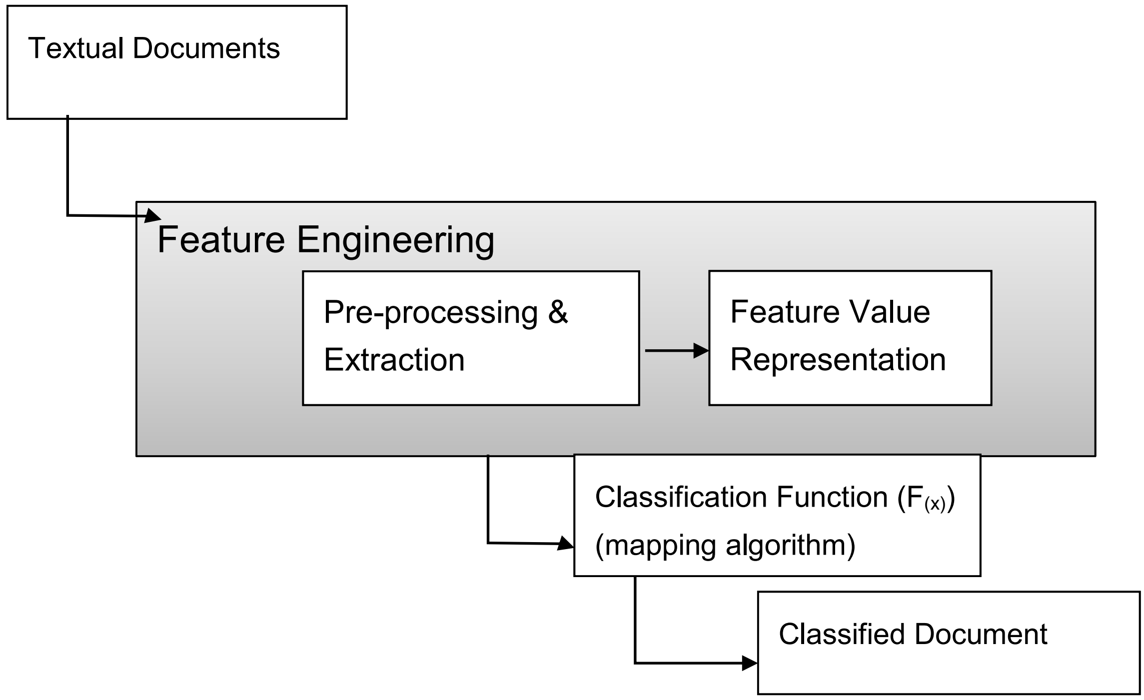


Figure 3.1 Text Classification process flow

The figure shows the process flow of a typical Text Classification process pipeline. A textual document goes through the feature engineering phase when the document is pre-processed, and features are extracted and represented in a suitable format for an algorithm, which is also referred to as the classification function f(x), to be applied. After this the document is assigned a given category based on the features extracted from the document and the pattern the algorithm was able to learn from those features, based on what it has been able to learn during training as described above. The feature engineering phase is discussed in detail in the next section.

3.3 Feature Engineering for Machine Learning

The aim of identifying appropriate features for a classification task is to achieve both computational efficiency and accuracy. This stage of the process has been an active area of research in its own right. It is considered a critical stage of the process and has resulted in a sub-discipline within the Text Classification research community known as *Feature Engineering* (Scott and Matwin, 1999). For example, reducing the features could lead to reduction in the feature space, which could result in high efficiency, but not necessarily lead to high accuracy. This is due to the fact that relevant features might not have been considered. Some of the features considered during Text Classification are discussed. As Figure 3.1 demonstrates, feature engineering encompasses two sub-processes: Feature pre-processing and extraction, and feature representation for Machine Learning. We discuss these sub-processes in turn.

3.3.1 Feature Pre-processing and Extraction

This section discusses some of the feature extraction approaches employed in Computational Linguistics in general, including Text Classification available in the literature. However the features discussed here are not exhaustive as there are other features employed in various tasks such as automatic genre identification (Sriurai et al 2010).

The Bag of Words approach

The Bag-of-words (BoW) is a common approach to modelling features and is regarded as being the baseline for evaluating Text Classification systems (Scott and Matwin, 1999, Sebastiani, 2002, Lamontagne et al., 2006, Moschitti and Basili, 2004). With this approach, each distinct word in the document is considered as a feature of that document. It must be pointed out here that the definition of what constitutes a word or term varies with respect to the application or domain of interest. This is a major decision that must be taken into account with this approach, and it makes up part of the pre-processing stage. Despite its popularity in the literature and the success stories about the BoW approach (Lewis, 1992), it has been noted to have various issues: the BoW approach does not make use of all available information in the original document, including the order of the words in the sentence. In other words, BoW by its very nature cannot take syntactic structure of a document into account. Another problem with this approach is its inability to recognise synonyms of a single word, consequently resulting in increased sparseness of the feature space (Wang and Domeniconi, 2008).

Phrase-based approach

The phrase-based approach seeks to select the phrases from content that best describe the document. This means words are expressed together to give contextual meaning, as they may not make sense contextually when considered individually. For example the words *health* and *worker* have their individual meaning, but will not make sense in this context until combined to form the phrase *health worker,*' which gives information that describes the occupation of an individual. This is based on the assumption that they contain useful information about the documents. The process therefore attempts to employ lexical and information retrieval techniques to extract phrases from the textual document that are most likely to describe it, and present them together as features of the document (Witten et al. 1999). Several linguistic techniques have been developed to identify phrases in the literature. Etzioni et al.(2005) for example employed a rule-based approach implementation using regular expressions, which is based on Part-of-Speech (PoS) tags obtained from the Brill tagger(Brill, 1995). Biber and Barbieri (2007) also proposed the use of lexical bundles to identify phrases and multi-words found in university spoken and written registries.

Ontology-based approach

Another approach to feature engineering for Text Classification is the use of some form of ontology resource. This enables relationships between terms and phrases to be captured (Agrawal and Kakde, 2013). As also pointed out by Spasic et al.(2005), ontologies tend to provide a mechanism that allows the appropriate sense of a word to be identified in a given text. This, again, is to overcome the problem of the BoW approach which tends to ignore the semantic relationships between words in a document. For example *bleed* has a relation *is a* with *symptom*. The identification of these relations can be achieved through the use of a knowledge base and resources such as ontologies or a thesaurus for the given domain. Furthermore, a corpus annotated with semantic relationships between concepts is another resource, for example the work carried out by Roberts et al (2009), which focused on producing an annotated corpus for the clinical domain.

Research carried out by several groups suggests improvements in performance when ontologies such as WordNet (Miller, 1995) are applied to capture the semantic relationships between words in the document for classification. For example, a recent experiment carried out by Wang and Domeniconi (2008) was reported to have achieved higher performance over BoW when they applied Wikipedia as a resource to extract relationships between words in a document. The difficulty with this approach however is the process of constructing the ontology, which is mostly manual and a time consuming process. There is also an issue about coverage as new words appear which makes it possible and useful only in domain settings with limited and controlled vocabulary.

3.3.2 Feature Value Representation

Feature value representation, also referred to as *Term Weighting* is a technique that allows some form of value or weight to be associated with the terms found in the document in order for a vector to be constructed. This is then presented to Machine Learning classifiers as earlier described. It is a technique proposed by Salton and Buckley (1988) to represent documents as vectors, popularly employed in information retrieval and more recently applied to Text Classification. As already indicated, a feature of a document could either be a word or a phrase in any other form used to identify the content of the document. Regardless of the scheme of representation, each feature must be associated with a value or weight, which indicates the importance of the feature in terms of its contribution to the classification. As argued by Leopold and Kindermann (2002) the weighting strategy employed has bigger implications for the accuracy of classification than the choice of learning algorithm employed in the classification process. There are numerous term weighting schemes proposed in the literature by various researchers(Liu et al., 2009, Lan et al., 2009, Lan et al., 2005). However, these weighting schemes are variants of the three basic and standard schemes as summarised below:

Table 2.1 Term weighting schemes

Scheme	Description
Binary	Boolean logic representation; 1 = present, 0 = not present
TF	Frequency count of terms found in a given document
DF	Frequency count of documents that contain a given term.

As noted by Jurafsky et al.(2000), the approach employed in assigning weights to tokens has an impact on the performance of the intended system in terms of precision and recall. This is due to the fact that the majority of classification algorithms apply a ‘strength of evidence’ policy during weight assessment and therefore binary values are not indicative enough of a token or a term in a given document. All weighting strategies are centred on two basic statistical principles as noted by Sebastiani (2002):

- the indicator of the importance of a term in a document, which is the frequency that term occurs in a given document (TF) and ;

- the more documents (DF) a term appears in, the less the term serves to semantically differentiate the documents.

With the exception of the binary approach, which represents feature occurrence as '1' and non-occurrence as '0', the other two approaches suggest weights based on frequency counts of either the feature or the documents containing the feature, based on the above mentioned basic statistical principles. While these schemes are sometimes employed as stand-alone, they are also sometimes mathematically combined. For example, the DF and TF are mostly combined by the product of the TF and the inverse of DF (iDF) to form another widely used scheme known as TFiDF (Salton and Buckley, 1988). The idea for this combination is that the higher the frequency of a term in a given document, the more it is a representative of its content. Also, the more documents a term occurs in, the less powerful it is in discriminating between a given set of documents (Wu et al., 2008). Recent advancement in research in this subfield has seen more sophisticated approaches; a combination of feature selection algorithms such as information gain, Chi-square, gain ratio and odd ratios with TF and DF have been explored (Lan et al., 2009, Liu et al., 2009). This has led to categorisation of term weighting schemes into supervised and unsupervised methods due to the process employed in estimating the values (Lan et al., 2005). Furthermore, DF or TF is sometimes combined with a normalisation factor. For example a normalised factor of document length takes into account terms of the same frequency in different documents to ensure features found in both short and long documents are treated with equal importance(Debole and Sebastiani, 2003).

3.4 Related Work: Biomedical Text Classification

The classification of biomedical documents has been witnessing a high rate of growth in research in the applications of Text Classification technology(Pakhomov et al., 2008, Pakhomov et al., 2007, Cohen and Hersh, 2005, Cohen, 2006). Cohen (2006) for example employed Chi-square as a statistical technique to extract features for a Support Vector Machine algorithm classifying genomes in biomedical text. Similarly, Pakhomov et al.(2008) employed various Text Classification based approaches to developing predictive models that identified patients with risk of heart failure from clinical notes obtained from Electronic Health Records.

The studies mentioned above have mainly explored data originating from the formal environmental settings of the biomedical domain, where use of language has been standardized with limited vocabulary. However, limited research has explored informal settings where there are no specific rules but rather colloquial language has been predominantly used. Nikfarjam and Gonzalez (2011) and Leaman et al.(2010) are among the few researchers who have explored colloquial text within the biomedical domain. For example, Nikfarjam and Gonzalez (2011) employed Computational Linguistics approaches for automatically classifying whether users experienced adverse reactions to a given drug. Using data generated from *DailyStrength* (www.dailystrength.org), they employed association rules to extract patterns of colloquial expressions that correlated with adverse reactions. Their work was largely motivated by works in the area of automatic analysis and classification of sentiments and opinions, which have mostly been expressed in colloquial text (Gamon, 2004, Oberlander and Nowson, 2006, Turney, 2002, Pang and Lee, 2005). Pang and Lee (2004) for example, employed Computational Linguistics approaches to determine whether a sentiment expressed about a movie was positive, negative or neutral. Using various lexical and statistical features derived from a sample of movie review text, they demonstrated the possibility of using this approach with comparable results to the one obtained by humans.

The above evidence suggests an emerging interest in research focused on automatic classification of colloquial text in general and specifically text from the biomedical domain. However, these efforts have not been extended to VA open narrative text, which is considered a rather unusual subtype of the biomedical genre as argued by Danso et al.(2013a).

3.5 Summary

This Chapter explored the processes and techniques employed in the classification of text in general. It described the processes involved and the logical principles that underpin Text Classification. The Chapter further provided a detailed description of how features are extracted and represented in order to harness the powers of a Machine Learning algorithm to learn patterns from these features. The Chapter concluded by conducting a survey into the applications of text classification within the biomedical domain and how these are related to classification of VA text.

The principle and techniques reviewed have successfully been applied to biomedical text from both formal and informal settings. However, these techniques have not been applied to VA, which is also considered a type of biomedical text obtained from an informal setting. This research is the first to explore how these techniques could be applied to VA as text originating from a different setting and application domain.

Chapter 4

Machine Learning Methods For Classification

"I'd rather learn from one bird how to sing than to teach ten thousand stars how not to dance" E.E. Cummings

4.1 Introduction

In Chapter Two we provided a definition of Machine Learning, whereby a computer algorithm can be said to have learned once it is demonstrated that the algorithm it uses has improved in its performance of a given task over a period of time through experience (Mitchell, 1997).We also made an attempt to formalise this definition. Numerous Machine Learning algorithms have been employed to tackle various classification problems (Kotsiantis et al., 2006). In Chapter Three we provided an overview of the Text Classification process, and the stage where a Machine Learning algorithm is employed. As Figure 3.1 suggested the selection and creation of a Machine Learning model is the next step once the document representation scheme is finalised.

In this Chapter we will discuss some of the Machine Learning algorithms that have successfully been employed in carrying out classification tasks: Naïve Bayes, Support Vector Machines, and Decision Trees. These algorithms could be classified into the three main families: probabilistic or generative, linear or discriminative and rule-based or decision trees (Reeves and Quigley, 1997). Additionally, some variants of these algorithms have previously been employed in predicting causes of death from the closed responses part of VA as previously discussed in Chapter Three. Our motivation is also based on the philosophy and the underlying assumptions employed, which vary between algorithms. In view of this, it seems natural to review these algorithms and apply them in our work to observe how they perform on the open narrative text and a combination of that and the closed response part of VAs.

4.2 Machine Learning Algorithms for Classification

4.2.1 Support Vector Machine

The Support Vector Machine (SVM) classification algorithm is one of the newer supervised Machine Learning techniques found in the literature, and has proven to be robust in dealing with noisy and sparse datasets. As a result, it has been the preferred technique to be applied to classification problems (Kotsiantis et al., 2006). SVM was originally proposed by *Vapnik* (1999) to deal with classification problems, and the principles under which SVM operates could be described as a hybrid of linear and non-linear, based on the *Structural Risk Minimisation* principle (Vapnik, 1999). The SVM formula is as defined below in equation (2):

$$\min_{\theta} C \sum_{i=1}^{m} \left[y^{(i)} cost_1(\theta^T x^{(i)}) + (1 - y^{(i)}) cost_0(\theta^T x^{(i)}) \right] + \frac{1}{2} \sum_{i=1}^{n} \theta_j^2$$

Where:

C is a parameter for the complexity factor to be tuned based on the characteristics of the underlying data and domain of interest; *y* is the category label, where *y*(positive) = 1; or y(negative) = 0; x is the input vector or feature input; n =number of features ; m= number of examples in training set; $cost_0$ is a similarity function for determining the category (y) given instance (i) when y = 0 and is $cost_1$ when y = 1; θ^T= transformed feature vector, which is the inner product of the support vectors. Support vectors are derived vectors obtained from the input vector, x.

During learning, SVM employs a technique of 'maximal-margin-hyper-plane' as part of the similarity function, such as cosine similarity measure (Charikar, 2002), to determine the decision boundary that separates category y= 1(positive) from category y = 0(negative) as shown in Figure 4.1

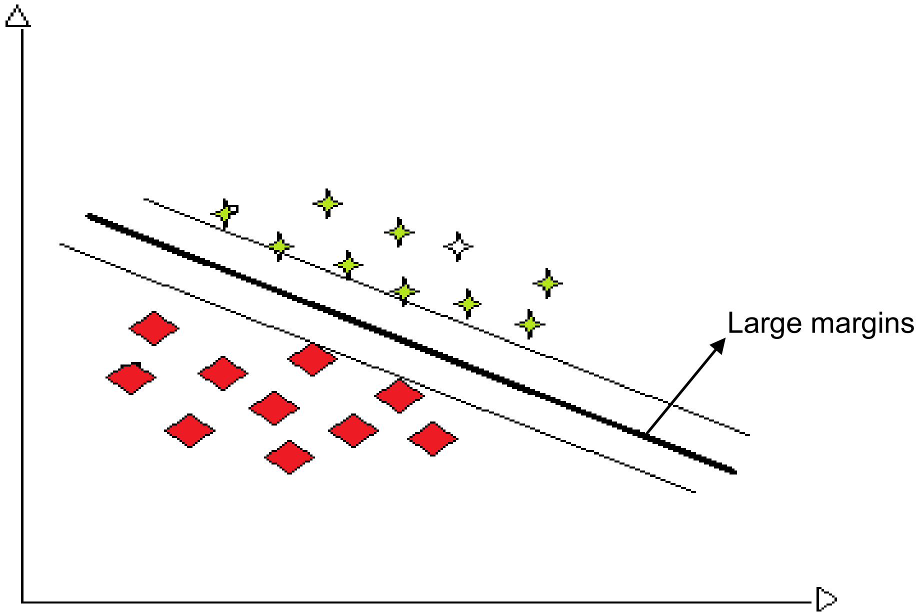


Figure 4.1 graphical representation of SVM learning algorithm showing red as category y (positive data points) and green as category y (negative data points)

The 'large margins' shows the hyper-plane that provides the maximum distance that linearly separates the categories. However, where this cannot be achieved because non-linearity exists, SVM has the ability to adapt by employing ‘kernels’ as part of similarity functions and is able to map the non-linearity that exists between category labels and feature space. The resulting hyper-plane established in the feature space by this kernel provides a direct mapping to the non-linearity structure that exists within the feature space (Kotsiantis et al., 2006). Despite its competencies discussed above, SVM tends to be computationally expensive by virtue of the kernel technique it employs during learning. This, however, can be minimized during SVM model training and evaluation since the kernel is a parameter that can be adjusted depending on performance, which eventually reduces computational cost.

Despite the success in the application of SVM to various problems, there has been continued research with the aim of improving its performance. This has resulted in variants of SVM. An example is the Sequential Minimisation Optimisation (SMO) algorithm developed by Platt (1999). SMO is considered to be faster and relatively easier to implement compared to the original SVM.

However, Keerthi et al.(2001) identified some deficiencies and proposed improvements to the SMO algorithm, which aim to address the problem of inefficiency in computing a threshold value that otherwise tends to prolong the internal computational process.

It is worth noting that SVM was originally designed for binary classification tasks, where one category represents data with positive examples (y=1) and the other represents negative examples (y=0) as described in the above equations. However, for a multi-class classification task, one-vs-all strategy is employed, where the cost function is optimised for each of *K* classes (Rifkin and Klautau, 2004), which is well suited for the classification problem under investigation and thus is applied in this research.

4.2.2 Naïve Bayes

Naïve Bayes (NB) is a generative-based algorithm which is considered to be relatively simple, it being based on probability models derived from the Bayesian theorem (Jiang et al., 2007). This classification technique analyses the relationship between each feature and the category for each instance to derive a conditional probability for the relationship between the feature values and the category. The conceptual framework for NB is based on joint probabilities of features and categories to estimate the probabilities of a given document belonging to a given category.

Like other probabilistic approaches, NB made the strong assumption about the data as demonstrated in Figure 4.2. This figure shows the process of learning, where NB assumes that all features of the training set examples are independent of each other given the context of the category, making it suitable for classification tasks with large number of features such as a Text Classification (McCallum and Nigam, 1998).

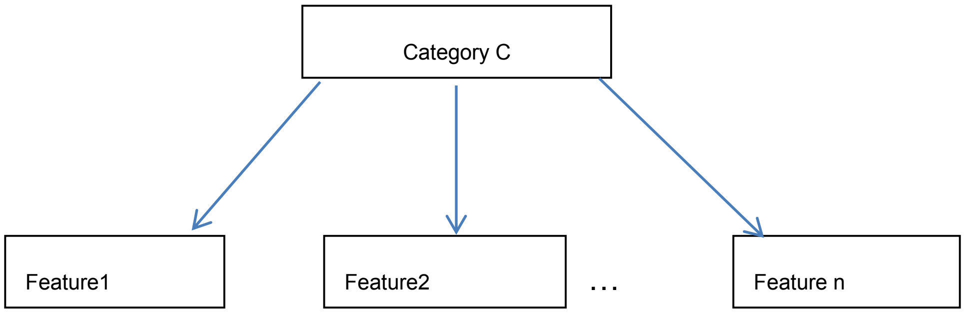


Figure 4.2 graphical representation of Naïve Bayes

It then posits a probabilistic model that embodies this assumption to estimate the parameters obtained from the training set to build a generative model during the learning process. NB has also proven to be robust to noise and missing data as it has the ability of generating relatively accurate models without having any impact on the final outcome (Witten and Frank, 2005). Its relative simplicity is also an indication of why it tends to be more efficient than the majority of the classification techniques found in the literature (Amor et al., 2004).

4.2.3 Decision Trees

Decision Trees (DT) belong to the rule-based family of Machine Learning algorithms, which have been successfully applied to different domains (Sebastiani, 2002). Despite the fact that it can be regarded as a relatively old technique, DT has stood the test of time. For example, DT has recently been employed as a Machine Learning technique to develop classification models that automatically classify pancreatic cancer data (Ge and Wong, 2008). DT based algorithms 'learn' from training examples by classifying instances and sorting them based on feature values. Each node in a DT represents a feature of an instance to be classified, and each branch represents a value that the node can include in making a decision. Figure 4.3 below shows an illustration of how DT works within the feature space.

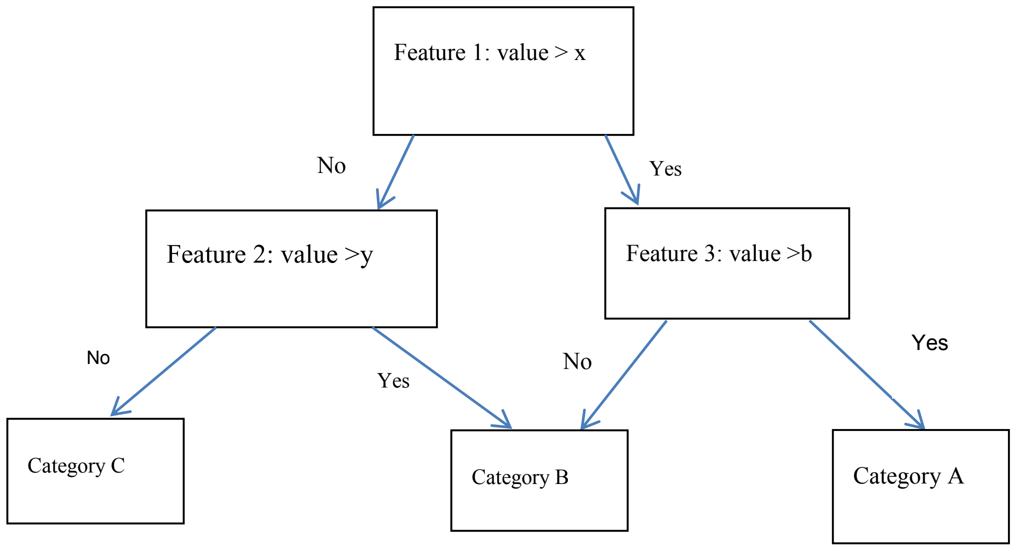


Figure 4.3 graphical representation of Decision Tree

The algorithm starts the process at a root node of the tree. This root node is established by finding the feature that best divides the feature space based on the feature values, and there are numerous approaches to identifying the best feature (Kotsiantis et al., 2006). Due to the approach DT uses to search for a solution within the problem space, efficiency tends to be an issue, especially when dealing with large datasets. This has resulted in research attempts into how this could be improved. Nevertheless, DT is characterised by its relatively transparent outputs, which are easy to be read and understood by humans. DT has been shown to have superior performance over other techniques with regard to some specific domains with datasets that have discrete/categorical data feature values (Anthony, 2005).

4.3 Validation Methods

Validation is the process of determining the performance accuracy of a Machine Learning algorithm after learning from a training set. There are various methods employed in carrying out validation and in this section we will discuss some of those methods.

4.3.1 Hold-out.

The hold-out method, which is also known as test sample estimation, is based on a split of the data into two mutually exclusive sub-sets called the training set and the test set. This division, which is usually 75% training set, is used for training the learning algorithm with the remaining 25% test set used for testing and estimating the performance of the algorithm (Kohavi, 1995). This method is computationally less expensive compared to the K-fold cross validation. There are however drawbacks associated with this method. As Kohavi (1995) points out, the accuracy of the estimation could be problematic. For example, the more examples that are retained for the test set the higher the bias of the performance estimator. On the contrary, the fewer the test cases the wider the confidence interval for the accuracy of the performance estimator. The holdout method also tends to underutilise (about 25% unused) the data, which could then instead have been used as part of the training.

4.3.2 K- fold Cross-validation

Cross validation is a well-established method of estimating performance of a Machine Learning algorithm. The nature of the procedure involved has resulted in another name known as the *rotation estimate* (Kohavi, 1995). It involves randomly splitting the training data into K-mutually exclusive subsets, where folds refers to the subsets and K is the number. The learning algorithm is trained and tested K times in turn for each fold and the overall accuracy is estimated. The formulae employed in the calculation will be discussed later. The choice of the value for K depends on the dataset. However, K= 10 has been found to be a good number of folds, which corresponds to 90% training and 10% testing (Pereira et al., 2009).

The problem, however, with the K-fold cross validation is that it does not take into account category distribution, which could be problematic for skewed or imbalanced datasets, where some classes are rare. To address this problem *stratified cross validation* has been introduced. This enables each stratified fold to contain approximately an equal proportion of category labels in the dataset (Kohavi, 1995).

K-fold cross validation tends to be useful when there is limited labelled training data. The averages obtained from the rounds of training and testing tend to provide good estimates for algorithm performance on an unseen or new dataset. Furthermore, K-fold cross validation has also been found to be robust to the variance and bias problems (described in Section 4.5) that tend to be associated with training data (Kohavi, 1995). The drawback, however, is that this could be relatively computationally expensive compared to other methods (Burman, 1989).

4.4 Evaluation metrics

The performance of Machine Learning algorithms can be measured by various approaches; the most common metrics being precision and recall. Figure 4.4 shows a contingency table used in estimating these performance measures (Lewis and Jones, 1996). The Figure provides a framework for evaluating the performance of a classification algorithm. It shows a matrix for determining the number of the classification distributions between a classifier and an expert judgment, which is referred to as the gold-standard or ground truth. These evaluation metrics being discussed will therefore make reference to this Figure.

		Expert Judgement	
		Yes	No
Classification Algorithm	Yes	True Positive (TP)	False Positive (FP)
	No	False Negative (FN)	True Negative (TN)

Figure 4.4 contingency table for classifier evaluation

4.4.1 Overall Accuracy

The overall accuracy of an algorithm can be determined based on the values obtained from figure 4.4 as expressed in an equation as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

4.4.2 Precision and Recall

Precision is a measure of the success rate of assignment of all documents correctly classified based on the categories assigned by the domain expert, which in this instance is the CoD certified by the physicians. This is expressed in an equation as:

$$\text{Precision (Pr)} = \frac{\text{TP}}{\text{TP+FP}}$$

Recall is the measure of success rate assignment of all documents that were supposed to have been assigned correctly by the algorithm based on the domain expert. This is expressed in an equation as:

$$\text{Recall (Re)} = \frac{\text{TP}}{\text{TP+FN}}$$

4.4.3 Single measure: F-measure

Apart from these basic methods for measuring the performance of learning algorithms, other approaches such as *F-Measure* have been proposed (Lewis, 1992). F-measure was said to have been introduced with the aim of having a single score to evaluate algorithm performance. This can be expressed in an equation as:

$$\text{F-measure:} \ = \frac{2^{*}\,\text{Pr}\ {}^{*}\,\text{Re}}{\text{Pr} + \text{Re}}$$

Other metrics are based on *category similarity* and *distance* measures proposed by Sun and Lim (2001).

4.4.4 Misclassification Measure

Another method used to evaluate the performance of a Machine Learning algorithm is the misclassification rate. One metric used to determine the rate of misclassification is the *False Positive Rate* (FPR) (Fawcett, 2006). This is expressed in an equation as:

$$\text{False Positive Rate (FPR)} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

4.4.5 Macro and Micro Averaging

It must be noted that Figure 4.4, the equations and the evaluation metrics described above are for two-classification problems. To be able to evaluate the performance of a classifier for a multi-class classification problem, the above performance measures must be computed for each category (Sokolova and Lapalme, 2009). Macro and Micro averaging measures are useful for multi-class evaluations where the classes number more than two. Micro-average is said to assign equal importance to each document whereas the Macro-average assigns equal weight to each category (C_i…. C_k) (Sun and Lim, 2001) . These are expressed mathematically in equations as:

Micro-average:

$$[\![\text{Precision}(\text{Pr}]\!]^{\mu}) = \frac{\sum_{i=1}^{k} \quad (\text{TP}_i)}{\sum_{i=1}^{k(\text{TP}_i\ + \text{FP}_i)}}$$

$$\text{Recall}\ (\text{Re}^{\mu}) = \frac{\sum_{i=1}^{k} \quad (\text{TP}_i)}{\sum_{i=1}^{k} \quad (\text{TP}_i\ + \text{FN}_i)}$$

Macro-average:

$$\text{Precision }(\mathrm{Pr}^{k}) = \frac{\sum_{i=1}^{k} \quad (\mathrm{Pr}_i)}{k}$$

$$\text{Recall}(\mathrm{Re}^{k}) = \frac{\sum_{i=1}^{k} \quad (\mathrm{Re}_i)}{k}$$

Where, k= number of classes

F-measure can then be computed for both macro and micro-averages based on the equation above. Similarly, overall average accuracy and misclassification rates can also be computed. These evaluation metrics will be employed to measure the performance of Machine Learning algorithms to be explored for the prediction of cause of death from VAs in this thesis.

4.5 Issues Associated with Machine Learning Based Approaches

The characteristics of data for developing supervised Machine Learning models tend to present various challenges for learning algorithms. We will discuss some of these challenges in this section.

4.5.1 Imbalance and Sparseness

Data is said to be imbalanced when there is an unequal distribution of labelled data among categories resulting in there being more instances for some categories and fewer instances for others. Classification algorithms tend to function based on the principle of Occam's razor, which means that the simplest hypothesis that best describes the data is to be generated and represented as a generalisation of the sample data (Akbani et al., 2004). Consequently, this results in most standard classification algorithms overestimating the majority category and underestimating the minority classes (Chawla et al., 2004). In response, various strategies have been proposed to solve the imbalance problem in Machine Learning at both algorithm and data levels. The data level suggestions deal with random oversampling of the minority classes and under-sampling of the majority category. At the algorithmic level, *cost sensitive learning* strategies include making adjustments to the cost associated with the various classes to compensate for the imbalance; making probability estimates at the tree leaf (for decision tree algorithms); and adjusting decision boundary thresholds (Chawla et al., 2004). These costs and adjustments can however be difficult to determine and are dependent on the business case, which is the driving factor that will determine the trade-off between correct classification and miss-classification of documents. For example, the cost of misclassifying a VA document must be known and factored into the modelling process. This makes the data level approach a relatively easy strategy to implement. However, there are issues with either the under-sampling or oversampling of the data level based approaches. Under-sampling can lead to potential important information loss, whereas oversampling can lead to over fitting (Li et al., 2009). Notwithstanding the problems associated with the cost sensitive learning approach, comparative studies have demonstrated that it tends to outperform the data level approach (Japkowicz and Stephen, 2002).

Data sparseness is used to describe a situation where a document represented as a vector for Machine Learning contains few actual feature values and almost all the remaining features are empty (Joachims, 1998). The natural consequence of this phenomenon is learning from a high dimensional space, which can be an expensive process and also lead to inaccurate results (Chawla et al., 2004). An approach which has been proposed and remains an active research area is feature selection, a process which aims to select features that are indicative of the various categories.

The phenomena of data imbalance and sparseness have been noted as a major problem in Text Classification, which has led to the concept of the *curse of dimensionality*. This is discussed in detail in section 4.7. and has been a focus for research (Han et al., 2006, Nigam et al., 2000). For example, Zheng et al. (2004) demonstrated the ineffectiveness of employing some existing approaches to feature selection on imbalanced data and proposed a framework that tends to combine features selected from both positive and negative categories.

4.5.2 Sources and Effects of Noise

Noise is considered in the context of Machine Learning as any form of defect which tends to decrease the quality of data, and thereby affects the learning process of Machine Learning algorithms. This, in turn, results in performance problems and inaccurate predictions (Atla et al., 2011).

There are two main sources of noise found in data: noise originating from features and noise from category labels in training data (Weiss, 1995).These two types of noise have different behaviours and impacts on the learning algorithm. Feature noise has the tendency to corrupt examples in the training set. For example, consider a feature vector that contains a binary representation {11111}. Assuming the representation changes to {10011} as a result of feature noise such as missing data or corrupted feature values , the feature noise would correspond to about 40% of the changes to the original representation, which would thus have the tendency to change the prediction or learning outcome of the classifier. This type of noise tends to cause examples in the majority category to over shadow the examples in the minority category, which can lead to feature overlaps and consequently cause higher misclassification rates.

Category label noise refers to errors occurring in labelling or annotation of training data. This could be problematic, especially in supervised classification tasks where models are developed based on the category labels under the assumption that they are accurate and well defined. However, as pointed out by Hand (2006) these errors do occur. Noise originating from category labels does not necessarily lead to misclassification. The tendency of misclassification is high if the noise is in a category that is particularly high, leading to the algorithm learning and attributing the learned example to a wrong category (Weiss, 1995).

4.5.3 The Effect of Bias and Variance

Bias and Variance are two concepts that are used to describe two related phenomena, and a combination of these two phenomena tends to play a key role in the performance of a learning algorithm. Bias is a term used to describe the systematic errors that are generated by a learning algorithm. Variance describes random behaviour of the learning algorithm that occurs as a result of random variations and noise in the training data (Kong and Dietterich, 1995). In other words, Variance can be used to determine the sensitivity of an algorithm to the training set. An algorithm is considered less sensitive to variance if it is more stable to variations in training data. Bias on the contrary determines on average how close an algorithm is able to estimate from the "initial hypothesis" set by the algorithm to a "true hypothesis" as found in the training data in order to obtain a generalisable model (Friedman, 1997).

A good solution to having a better performing algorithm is to have both Variance and bias as low as possible. However, as indicated, because bias is related to the underlying design principles of the algorithm, it is important to take into account the selection of an algorithm for a given task. For example, Naïve-Bayes has been noted to have the ability to handle datasets that are characterised with a high degree of bias associated with them (Friedman, 1997), while the Support Vector Machine is an example of a relatively low bias or unbiased algorithm (Valentini and Dietterich, 2004). In contrast, because Variance is associated with the degree of variations found in the training data, one possible solution which has been proposed is increasing the size of the training examples (Friedman, 1997). The concept of *Ensemble Learning* has been proposed as a possible approach to reducing bias and variance in Machine Learning (Valentini and Dietterich, 2004). Ensemble-based learning encompasses the concept of combining algorithms to allow various hypotheses to be generated for the various individual algorithms based on their properties and configurations. A voting process is subsequently employed to determine the best hypothesis and generalisability capabilities for the training data.

4.6 The Curse of High Dimensionality

Data sparseness potentially leads to high dimensionality. The *curse of dimensionality* is a term introduced to suggest the importance of filtering as part of the modelling process(Bellman, 1957) and has subsequently been explored in the literature (Kuo and Sloan, 2005). Filtering is part of the pre-processing stage, where features that have no use in differentiating between documents are ignored or eliminated. This could be commonly used words that run through all documents and have no statistically significant or discriminative powers to distinguish been documents. Examples include function words like articles *a*, conjunctions *and*, and prepositions *in*. The main purpose of this is not only to increase system performance such as speed of processing by reducing the space required by the application and the processing resources needed, but also performance accuracy of the algorithm. This view is supported by an experiment performed by Lewis and Jones (1996), which suggested that a dataset with high dimensionality tended to have a negative impact on the performance.

Apart from the stop words, various approaches have also been developed to help to evaluate and obtain information in terms of the contribution of each word or term in a document. This includes information gain (IG), Chi-square (x^2) and others as noted by both Hotho et al. (2005) and Forman (2003). A number of feature reduction methods and their effects on performance are explored and discussed in Chapter Seven of this thesis.

4.7 Summary

This Chapter explored the landscape of Machine Learning methods for carrying out classification tasks. It began with a description of the theoretical and the philosophical underpinnings that differentiate some selected Machine Learning algorithms employed in classification. We then further surveyed the various evaluation metrics available to evaluate the performance of a given algorithm. It was noted that consideration must be given to multi-class and imbalanced data during performance evaluation. For example, macro-average should be the preferred choice of measure when dealing with imbalanced data since this tends to account for differences in category distribution better than micro-average. The Chapter further looked at some of the issues associated with Machine Learning based approaches and how those issues tend to impact on performance. The issues were general and applicable to every domain and will thus be taken into account during model development for cause of death prediction in VAs.

Chapter 5

Surveillance Data Management and The Verbal Autopsy Corpus

"Maybe stories are just data with soul", Brene Brown

5.1 Introduction

In Chapter One, we indicated the need for a robust registration or surveillance system in order to be able to identify deaths and eventually the cause of the deaths. However, a robust surveillance system also requires robust data management to ensure the smooth running of the surveillance system. In this Chapter, we will describe the data management system deployed that supported such a surveillance system and enabled the identification of deaths for VA to be conducted. We will then further describe the methods employed in building the semantically annotated corpus of VA documents that were used for this thesis.

It must be noted that some of activities described in this Chapter were carried out during data collection as part of the large-scale epidemiological study known as the ObaapaVitA study (Kirkwood et al., 2010; Edmond et al., 2008; Hurt et al., 2013;Kirkwood et al., 2013). Briefly, the ObaapaVita study was led by Professor Betty Kirkwood of the London School of Hygiene and Tropical Medicine. It was a 10-year multi-million pound project jointly funded by the United Kingdom Department for International Development (DfID-UK), and United State Agency for International Development (USAID) to evaluate the effect of weekly supplementation of vitamin A to women of child-bearing age (15 - 45 years) on maternal mortality. This led to the establishment of a surveillance system which was used from December 2000, which facilitated the conducting of follow-on large scale epidemiological studies including the NewHints study (Kirkwood et al.,2013). Figure 5.1 shows a map of the area covered by the surveillance system which constituted seven contiguous districts, covering an area of over 240,000 Km^2 predominantly rural land within the Brong-Ahafo region of Ghana.

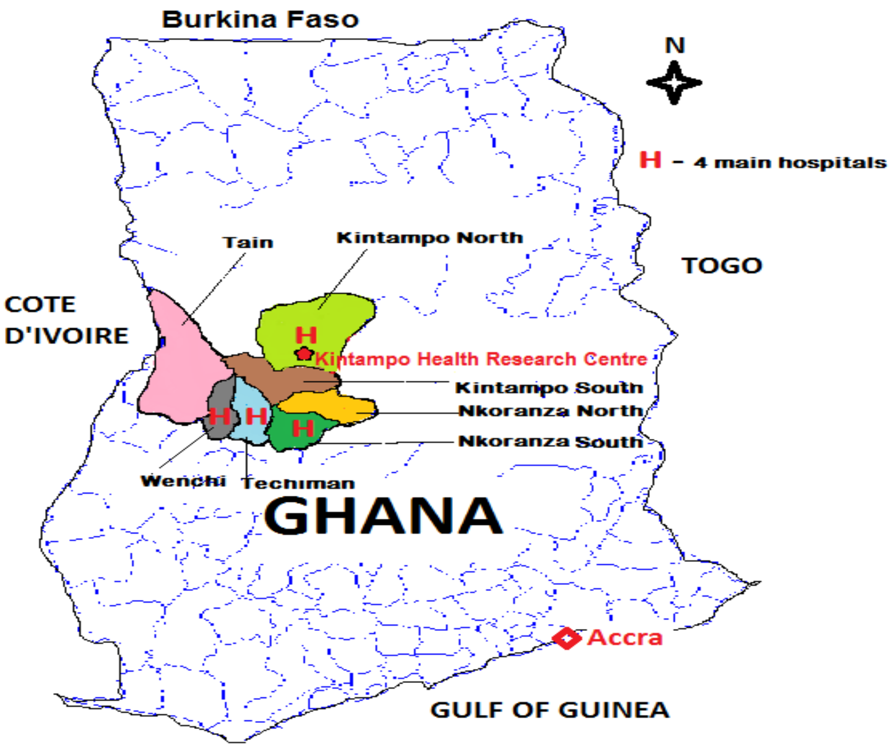


Figure 5.1 map of Ghana showing the seven districts from which the corpus was generated. Adapted from the original source obtained from the Geography Department, University of Ghana

Prior to the commencement of this PhD the author was a key member of the ObaapaVitA study team. The author was instrumental in all activities and was also responsible for some, especially those specific to this research, which are discussed further in this Chapter.

The Chapter concludes by carrying out analysis of the cause of death as obtained from the annotation process and proposes three re-grouping schemes that would be employed when carrying out our experiments. The rationale and the effects for the re-groupings are also discussed. To differentiate between the original label, all causes of death that are derived from our proposed scheme are underscored (_) as part of the labelling.

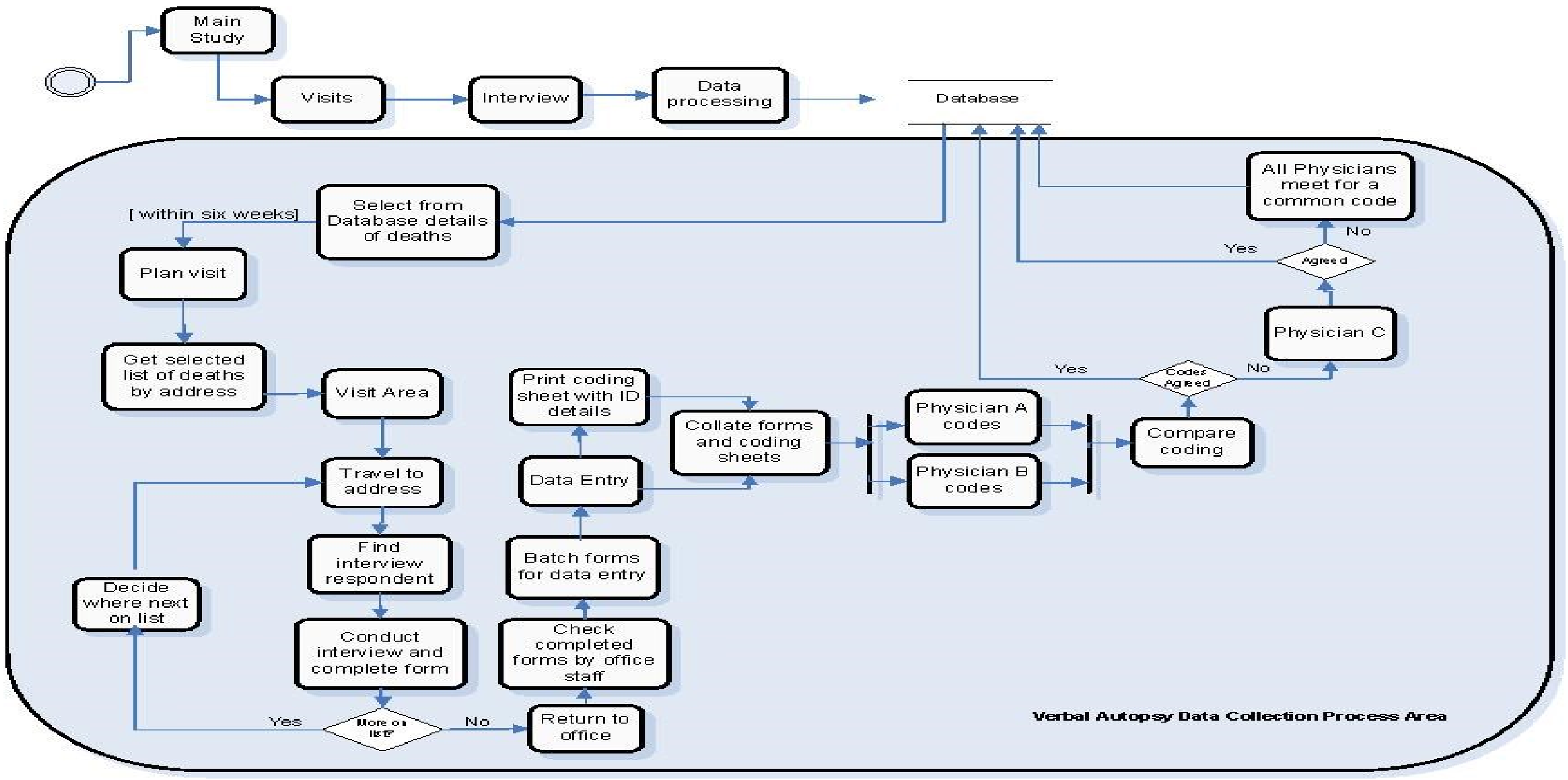


Figure 5.1 process flow of the ObaapaVitA surveillance system

5.2 Data Management

Figure 5.1 shows an outline of the process flow of the surveillance system deployed in Ghana which enabled the VA corpus used in this thesis to be captured.

The basic flow of activities illustrated in Figure 5.1 is summarised briefly. Routine visits were made to participants of the study for data to be collected. The data, which included events that had occurred concerning the study participants (including deaths) at the time of visit were processed and stored in a database. A list of deaths was then generated from the database which triggered the VA corpus collection process to begin. Interviewers planned visits according to the list generated from the database indicating which address to visit. An Interviewer travelled to the address and identified an appropriate respondent according to the protocol, which suggested a close friend, relative or other caregiver of the deceased, deemed capable of providing accurate information. The interview was conducted and a questionnaire filled in. The Interviewer then decided to move to another address for another VA interview or get back to the office. Quality control checks (e.g. no missing information) were performed on the questionnaires by office-based personnel and then they were forwarded for data entry in batches. Any issues such as missing or incomplete information identified at this stage would need to be resolved by the field team before forms were sent to the data processing department for processing.

The above process involved over 200,000 participants (women), and required an efficient and custom made Data Management System (DMS) to effectively manage the inherent complexities. The primary function of a DMS is to facilitate core tasks which are performed by a user: data quality control (cleaning, inconsistency and validation checks); store, retrieve and update data (Dodd, 1969). However, the design of any DMS is mainly driven by the functional and non-functional requirements imposed on the system by the users (Glinz, 2007). For example, a DMS developed for retrieval of information may require an efficient indexing algorithm to support that process. However, an updating process for example may not require the same level of complexity in design. The DMS developed for the surveillance system mentioned above required functionalities similar to an information retrieval system: it required an efficient retrieval of data for report generation; to support ad-hoc query requests; and also support batch processing of weekly updates and storage.

A detailed description of the DMS system can be found in the manual created by the author with input from the DMS team, which can be found in appendix B.

5.3 Building the Verbal Autopsy Corpus

5.3.1 Corpus Source and Sampling

This corpus was obtained from the surveillance system described above. The sample contained all stillbirths and deaths in infants to the age of 12 months, which is referred to in this thesis as the *infant sub-corpus*. Additionally, it also contains text about the cause of all deaths in adult women between the age of 15 and 45, which is referred to in this thesis as the *women sub-corpus*. The corpus contains a total of approximately 2.5 million words in 11,741 documents. Table 5.1 shows the breakdown of the corpus into the two sub corpuses.

Table 5.1 Basic statistics of the Verbal Autopsy corpus

Sub corpus	Number of documents	Number of Cause of Death Categories	Number of words
Infant	8,212	23	1.5 million
Women	3,529	43	1 million

The infant sub-corpus contains 8212 documents which is 1.5 million words with 23 categories. The women sub-corpus also contains 3529 documents accounting for 1million words of the total corpus with 43 causes of death categories.

5.3.2 The Interview Questionnaire

This is a standard questionnaire originally developed by the WHO and adapted by Edmond et al (2008) to collect the VA data as part of a study which aimed to validate its diagnostic accuracy within the study area shown in Figure 5.2. Edmond is a physician and domain expert, and was responsible for the adaptation and ensuring that all relevant details and symptoms of diseases were captured. The author of this thesis was responsible for the management and quality control of the data capture. The questionnaire includes identification, closed response and open narrative text sections.

Identification

This part of the questionnaire contained basic contact details of the deceased: Identification number; house number; name of subject; and date of birth.

Closed response part

The closed questions have in total over 200 variables for the infants and just over 270 for the women questionnaire. These variables are questions to elicit the presence of specific symptoms during the final illness. This information is often accompanied by a box for recording the length of time that the deceased experienced those symptoms. For example question number 6.1.8 in the Figure 2.1, asking a mother whether she had high blood pressure during the pregnancy. Where questions are not applicable, a double line is drawn through those questions and *99* is entered into the databases as the response value to differentiate it from missing information.

Open narrative text part

The purpose of the open narrative text is to enrich the data collected in the closed response as both the open narrative text and closed responses are used by physicians when attempting to determine the cause of death. As Figure 2.2 shows, the open narrative text part gives an interpretation of the narration of the event from the respondent by the data collector. It is important to emphasise that the narrative text captured is an interpretation of the interview, which is conducted in a local language (Twi) and the interviewer in turn interprets and summarises into English. This process is in parallel with the process of recording answers onto the paper questionnaire, with both being done in the course of the interview by the interviewer. It is therefore important that the information captured gives a true account of what transpired at the interview to enable accurate diagnosis to subsequently be carried out by physicians. Any inaccuracy in the interview process may result in an inaccurate diagnosis and the wrong cause of death being assigned. This therefore requires a high level of skills and experience in conducting interviews of this nature.

One notable difference between the women and the infant sub corpora that is worth pointing out here is that the structure of the questionnaire used for the infants was different from that used for the women. The narrative text part of the infant questionnaire had clearly defined sections that asked for specific information as shown in Appendix B. The women questionnaire however did not have sections but had blank pages that required the interviewer to hand-write during the interview. It is also important to note that the VA questionnaires were designed based on the most standard methods available at the time of the design. The adult women questionnaire was based on a validated instrument for adult deaths by Chandramohan et al.(1994) and an instrument for maternal deaths developed for a previous study by Campbell and Gipson (1993).

5.3.3 The Selection and Interview Process

As indicated, the corpus was generated as part of field trials, where routine visits were made to participants of the study for data to be collected. The data, which included events that occurred concerning the study participants (including deaths) at the time of visit were processed and stored in a database. A list of infant and adult female deaths was generated from the database to indicate which interviews needed to be conducted. This list was shared among a group of interviewers who had been trained to conduct VA interviews.

An earlier study suggested that a period between one and 12 months after death is recommended to elicit reliable information from respondents (Soleman et al., 2006). Although it is probably better to conduct the interview as soon as possible after the death to avoid recall difficulties, the researchers also have to respect the family’s need to grieve and a minimum of six weeks was therefore set before conducting the interview.

Background of the interviewers

The interviewers recruited for the VA information collection were people who had completed high school, spoke the local language as well as English, and had attained fieldwork supervisory status with no medical training. These people were recruited and provided with basic training on how to administer the VA questionnaire. For example, interviewers were given specific training on how to probe in order to elicit the information relevant to enable physicians to arrive at the possible cause of death. There was a mixture of both male and female interviewers and this did not seem to influence the data collection process.

The training of interviewers was carried out jointly by the lead author of Edmond et al (2008), Professor Karen Edmond and the author of this thesis. The training covered both questionnaire administration and the quality control procedures. This was to ensure data collectors understood both the questionnaire administration and the potential problems that could arise as a result of poor quality data.

5.3.4 The Annotation Process

For this research, annotation is defined as the process of reviewing the VA document and assigning a cause of death to it. We employed methods similar to the one described by Pestian et al. (2007) as a democratic principle approach to creating the final cause of death code for each VA document. The annotation process was jointly managed by the author and Karen Edmond (Edmond et al, 2008). I was responsible for the management, quality assurance processes and storage of the annotations assigned by the physicians. To achieve this, custom made software was developed to automatically identify differentials between annotations of a given VA document by two independent annotators as described above. Karen was responsible for the coordination of the activities of the process. The infant sub-corpus annotation scheme was adapted from the Neonatal and Intra-uterine Death classification to Etiology - NICE (Winbo et al., 1998) and the WHO Neonatal Child Health Epidemiological Reference Group- CHERG (Lawn et al., 2006). The ICD-10 (1992) classification scheme was adopted for the women sub-corpus. However emphasis was placed on the expected public health importance of the causes of death within the context of low-resource settings. Thus, where specific cause of death could not be determined due to insufficient information, the period in which the death occurred was assigned as the death outcome, referred to as *Time-of-Death*, which was a broad classification scheme that allowed deaths to be categorised. This was driven by the fact that having information about the period in which deaths occur is vitally important and more reliable than being without any information within the context of VA. It is also particularly useful for epidemiological studies (Quigley, 2005).

A coding sheet was generated with ID details of each death for use by the physicians during their review. The coding sheets and VA document were duplicated and a set passed to two different physicians. Each physician independently reviewed the VA document and recorded their judgement as to the most likely cause of death on the coding sheet. The sets of documents were returned and the coding sheets input into a database where a comparison between the two assigned causes of death was made using custom made software. An agreed code was assigned when the two physicians agreed. Otherwise, the process was repeated with a third physician. When there was an agreement between any two of the three assigned causes of death, then that was accepted; but when there was no agreement, a meeting was held between the physicians involved with the aim of an agreement being reached. In the event where there was still no agreement, the cause of death was assigned as *Unexplained* or *Cause Unknown or Not Ascertained* cause of death.

Alternative annotation process management models were explored to determine which would be least resource intensive without compromising on quality. The first model tried was that where annotators were given basic training in the annotation task; then they were given the completed questionnaires to review and produce the annotations at their own time and preferred location. In a second model, employed later, annotators were assembled at a central location to carry out the annotations.

Background of annotators.

Medically trained professionals were employed to carry out the annotations. The minimum qualification required to as an annotator is must have completed medical school and one year post qualification experience in a hospital. To ensure high quality annotation and a high level of agreement, all annotators were trained using standard annotation guidelines developed by the project management team and led by paediatricians for the infants and maternal health experts for the women.

As indicated earlier, the annotation approaches employed in this project also had options which could have eased the logistics burden imposed on the management of the process. One of the key challenges associated with the first approach was the difficulty in coordinating the activities to ensure that the processes were conducted in a timely manner. The second model however was relatively less difficult with regard to the coordination of activities. It was however logistically intensive and expensive in terms of the organisation. One approach that could possibly overcome these challenges is the adaptation of emerging annotation and text analytics tools such as GATE (Bontcheva et al., 2010). These tools offer the flexibility to allow annotators to work from any location and coordination can also be carried out with ease. It however requires an Internet connection, which has only recently become available in the developing world.

5.3.5 Corpus Transcription

We define transcription as the process of converting the written open narrative text into machine readable format. As seen from Figure 2.2 which demonstrates VA open narrative information captured the VA questionnaire requiring transcription. This process was planned and carried out by the author of this thesis. The author was responsible for all the activities described in this process.

Having recognized the need to transcribe the open narratives text into a machine-readable format, various options were explored to ensure the feasibility of its implementation, its usability and a justification for the chosen platform. A tailor made software tool was designed for this activity. The software had inbuilt spellchecking functionality that allowed typographical errors to be corrected. The next section of this Chapter gives a detailed description of the software.

The software

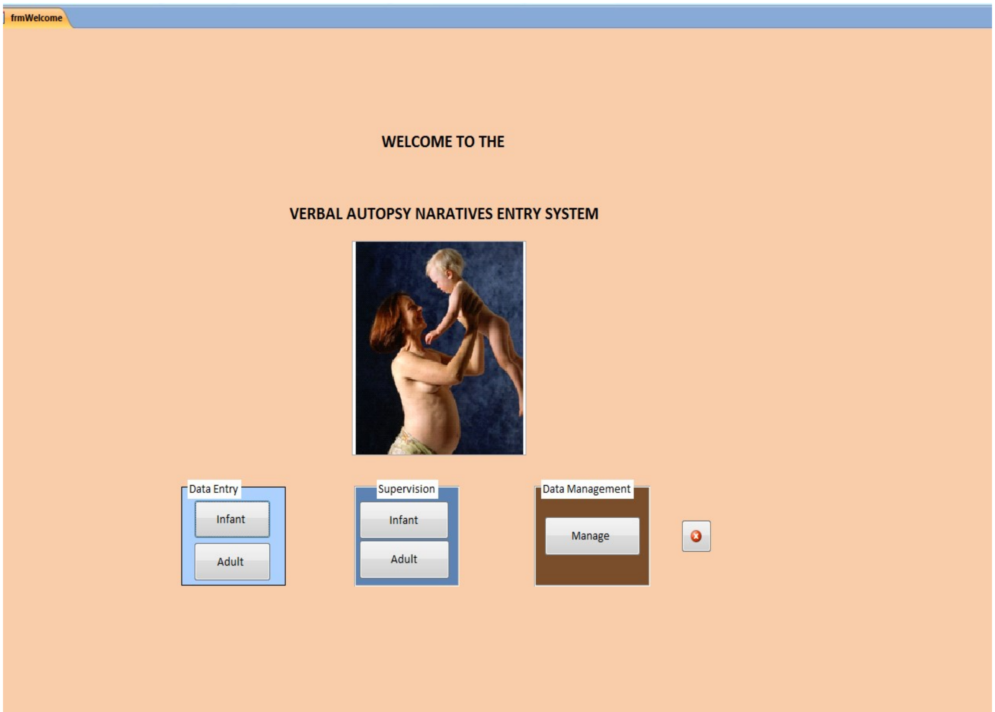


Figure 5.3 A ‘welcome’ screen of the system-the ‘gateway’ to all parts of the system.

The transcription software for capturingVA corpus data has been developed on Microsoft Access 2007 platform and has the following functionalities. It has a 2-tier architecture, having a flat front-end with basic business rules and a back-end used in storing the data. The system has three main modules; Data Entry, Supervision and Quality Control and Security. Functionalities available to each of these modules are discussed below.

Data Entry

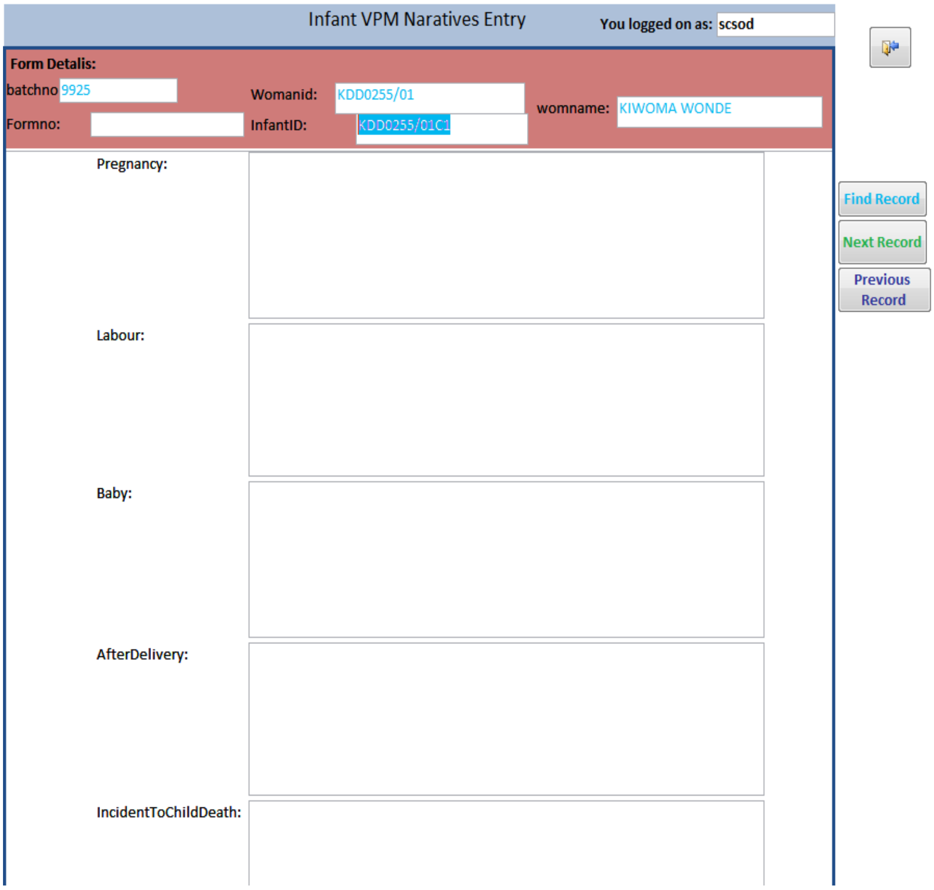


Figure 5.4 data entry screen of the transcription software

Figure 5.4 shows the data entry form on the main screen used for entering the narratives text. It populates the form with details of the individuals automatically from the database. The user is then given the option to locate an individual using the 'find record' button and specifying the ID of the individual for the system to retrieve from the database. This user can save a record after entry by clicking on 'Save record' button.

Supervision Module

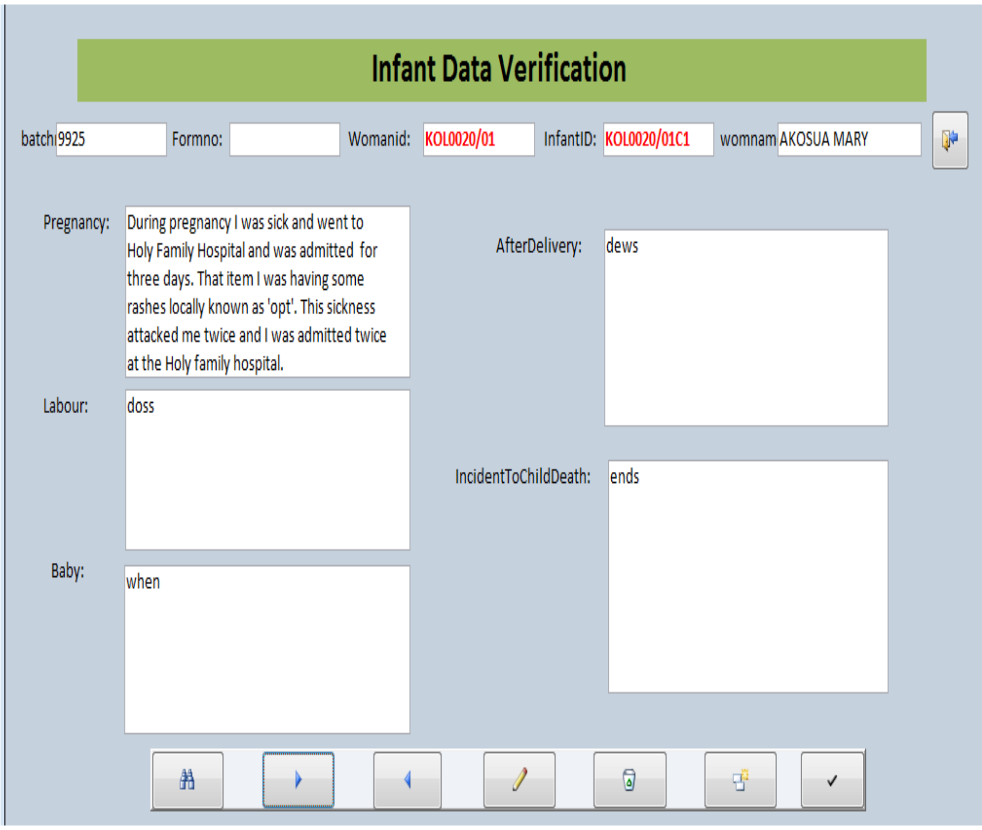


Figure 5.5 supervision module showing text to be verified

Figure 5.5 is the Supervision module which offers the opportunity for corrections to be made to the already entered narrative data. This is the next step after data entry. It has all the functionalities of the data entry screen in addition to data manipulating functionalities. Examples include *delete*, *duplicate finder*, and *add new record* buttons that are not present on the basic data entry screen.

Spell-checking

One of the requirements of the system is to have a spell-check functionality. This is an essential requirement, as misspelling could create potential parsing problems, which in turn could lead to the introduction of potential errors. The effect of this can be enormous as a study conducted by Ruch et al. (2003) demonstrated that 4% of misspellings of words translates into a 10% error at the sentence level.

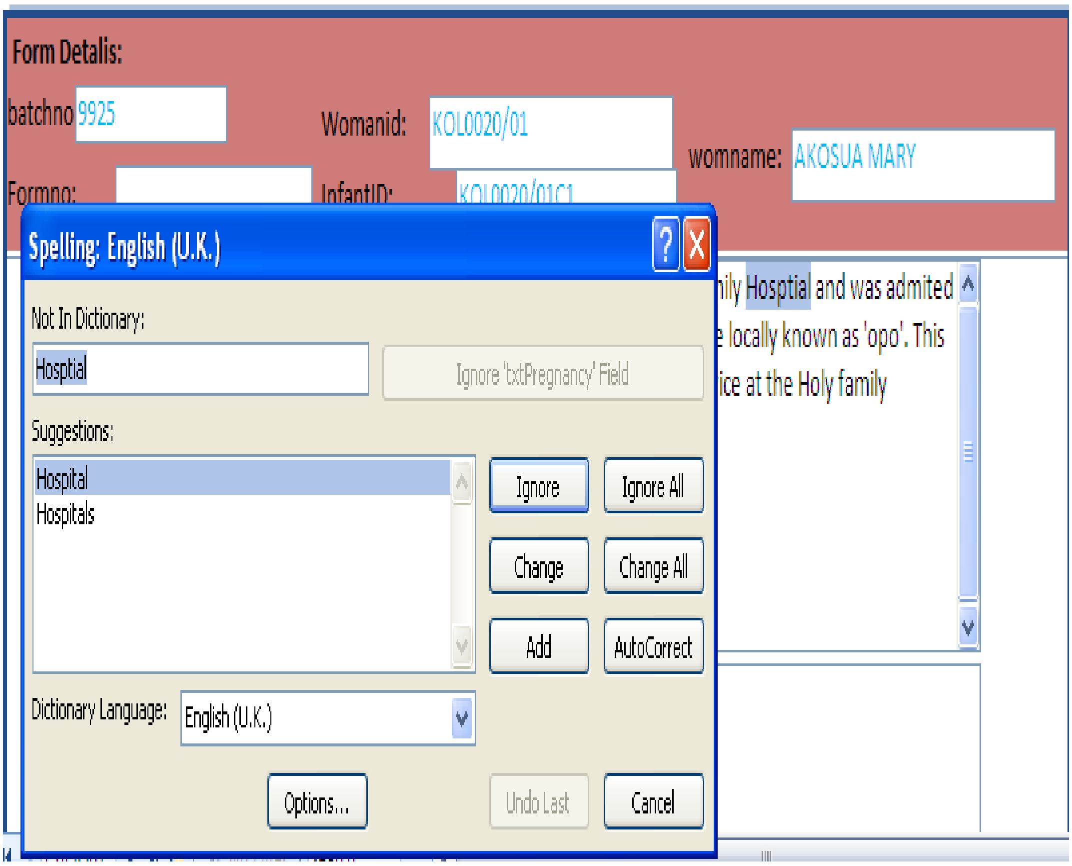


Figure 5.6 spell-checker

Figure 6.6 shows the spell-check functionality implemented, as part of the software identifying misspellings of words finds *hospital* in the pregnancy section of infant screen and offers suggestions of an appropriate spelling to the user. This functionality is used at the verification stage where Supervisors use it to verify or check on spelling errors the occurred during data entry. It must be noted that the spell-checker adapted for software is based on the generic one implemented as part of a general purpose word-processing application, and was therefore efficient in handling English words. However this meant that any VA specific words used had to be accepted by the user, which in this case meant the Supervisor had to update the dictionary.

Quality control and data security

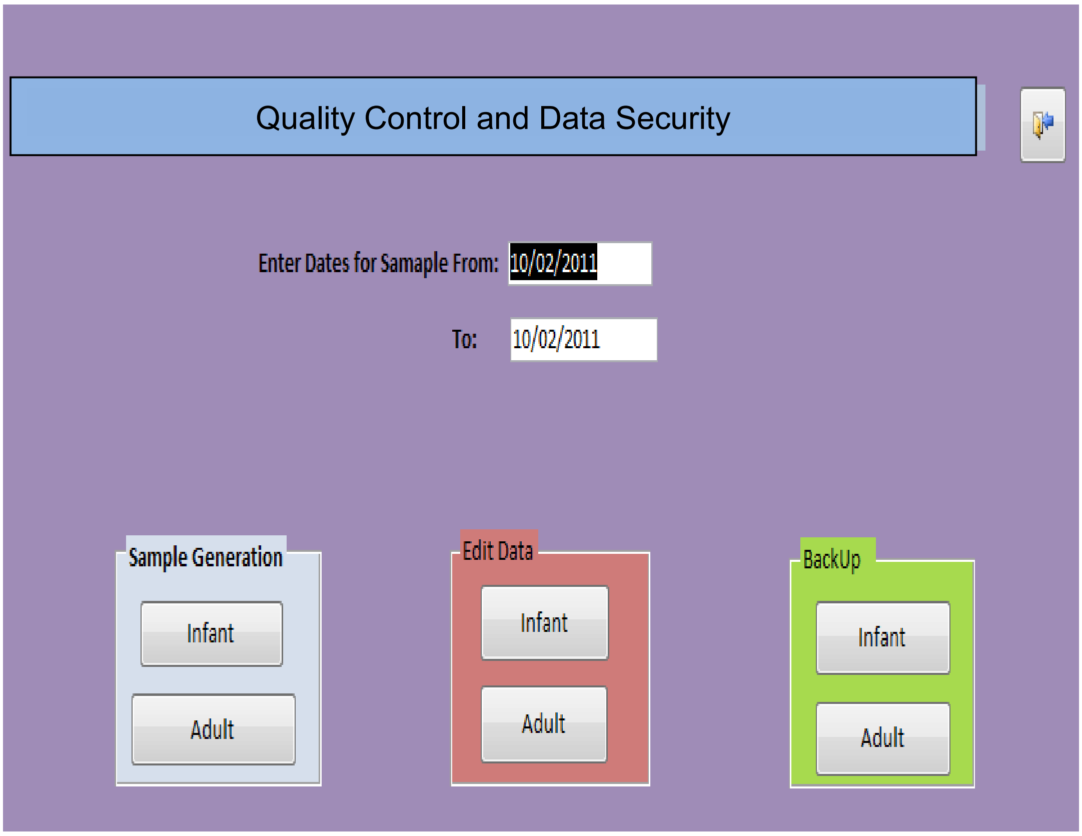


Figure 5.7 quality control and data security screen

Figure 5.7 shows the screen shot and the functionalities available for Quality Control and Data Security. This is the final stage of the process where a sample of the processed data is generated and cross-checked against the original forms. This is to ensure that the verification did not miss any errors created during data entry. A backup of the entire database is performed after verification has been completed on the samples using the backup functionality. The 'generate sample' checks for records which have been processed within a specified period (supposed to be after verification stage as per the procedure) and generates 10% of these records to be checked. Corrections can then be made to the original data where necessary by using the *Edit* button. Finally, the *backup* button allows a complete backup of the entire database to be performed to avoid tables being corrupted or any unforeseen eventualities.

The method employed in transcribing the corpus was one among several approaches that could have been explored, for example Optical Character Recognition or voice-to-text generation. Another approach could have been to electronically capture the corpus at the point of the interview by use of portable devices. With hindsight these approaches could have been more cost effective but the operational difficulties of the environment were challenging and a paper data collection approach with data entry clerks transcribing the results was a pragmatic choice.

5.3.6 Processing and storage

The software tool used for the transcription stored the corpus in a Microsoft Access database. This format was not directly usable for corpus analysis although the corpus is now available in a digital format. Also, the cause of death codes assigned by the physician were stored in a separate database and there was the need to have all this information stored a single file. The two separate databases were linked together using the unique identifier, which was assigned at the point of registering the study participant. The resulting database was then exported and encoded into an XML format as that format allows all the individual files to be stored as a single file. An example of this is shown in Figure 5.8.

```
<Cleaned_InfantVA_InfantVA_CoD>
        <infantid>
        KA017/1/08C1
        </infantid>
        <Pregnancy>
        I made eleven consecutive antenatal check-ups at birth xxx hospital and xxx maternity home respectively. I received two tetanus
        injections during the pregnancy and these, were the problems encountered : - painless bleeding at the initial stage of 2-3 months old
        (which I had drugs from xxx hospital for treatment), reported to have an anaemia at the age of 4 months old with headache/dizziness at
        the same time which, I had drugs and was advised to eat green vegetables. Besides these problems I had malaria (at the age of 3
        months), severe abdominal pains (6 months old) restlessness after meals of fatal movement at the left rib side at my seventh month. My
        pregnancy ended, at 11 months old at prince of peace maternity home through normal birth.
        </Pregnancy>
        <Labour>
        Labour started around 6:00am and gave birth around 1:00pm but had water (green in colour with no bad scent) breathing during labour.
        Problems had during labour were: - lower abdominal pains and the umbilical cord which came out before the child followed
        </Labour>
        <Baby>
        At birth, the child was an average in size with no malformation, but was having difficulty in breathing with blood flowing through the
        nose.
        </Baby>
        <AfterDelivery>
        At birth the blood flowing the nose was cleared, and artificial air was poured into the nose to assist him to breath well.
        </AfterDelivery>
        <IncidentToChildDeath>
        Child was having difficult breathing at birth and blood flowing through the nose as well. After clearing the blood from the nose an artificial
        air was pumped into the nose to assist him to breath well. The breathing rate was very fast and was abnormal. Even the breathing rate
        was very fast but at times, the breathing could stop for some time before it could come out again and the problems continued up to
        12:00pm (mid-day) and finally died at the maternity home. Another symptoms detected on the child was the physical colour changed
        into green at birth.
        </IncidentToChildDeath>
        <CoD>
        Neonatal-other causes
        </CoD>
<Cleaned_InfantVA_InfantVA_CoD>
```


Figure 5.8 XML tags mapping to various sections of infant VA document with cause of death information merged

5.3.7 Anonymisation

As with other clinical data and text in particular, the identity of subjects in a VA must be kept anonymous. However, due to time constraints on the project a detailed anonymisation could not be done on the open narrative free text, which included occasional named entity mentions such as persons and places. A complete anonymisation will be carried out before making the corpus publicly available. Nevertheless, partial anonymisation was carried out. To achieve this, basic details of all subjects were stored in a separate database and references were made to all documents using an ID which was allocated during the registration phase. The basic details database was obscured and was only accessible to the data collectors and trial management team. The challenge however was to find ways to deal with the named entity mentions and references made to individual details in the text. This challenge required careful handling since removing all the mentions and references in the text could lead to loss of potentially useful information. It was however found that the sensitive information, which required anonymisation found in the free text, could not mean much without the individual details. Even though there were occasional mentions of names of relatives these named entities could not be traced without additional information such as the location and address details which had already been removed. Pestian et al. (2007) made a similar observation in a corpus of clinical text, which was built for the development and evaluation of a multi-label classification shared task.

5.3.8 Ethical approval

Ethical approval was obtained from the relevant ethical committees to carry out the field studies and also this research. Consent was sought from the participants of the studies to use their data for research purposes. This however does not cover making the data publicly available and ethical approval for this purpose is therefore required.

5.4 Analysis of cause of death annotations

As previously mentioned the annotation scheme for infant sub-corpus adapted the Neonatal and Intrauterine Death classification to Etiology - NICE (Winbo et al. 1998) and the WHO Neonatal Child Health Epidemiological Reference - CHERG (Lawn et al., 2006). The women sub-corpus also adapted the ICD-10(1999) classification scheme.

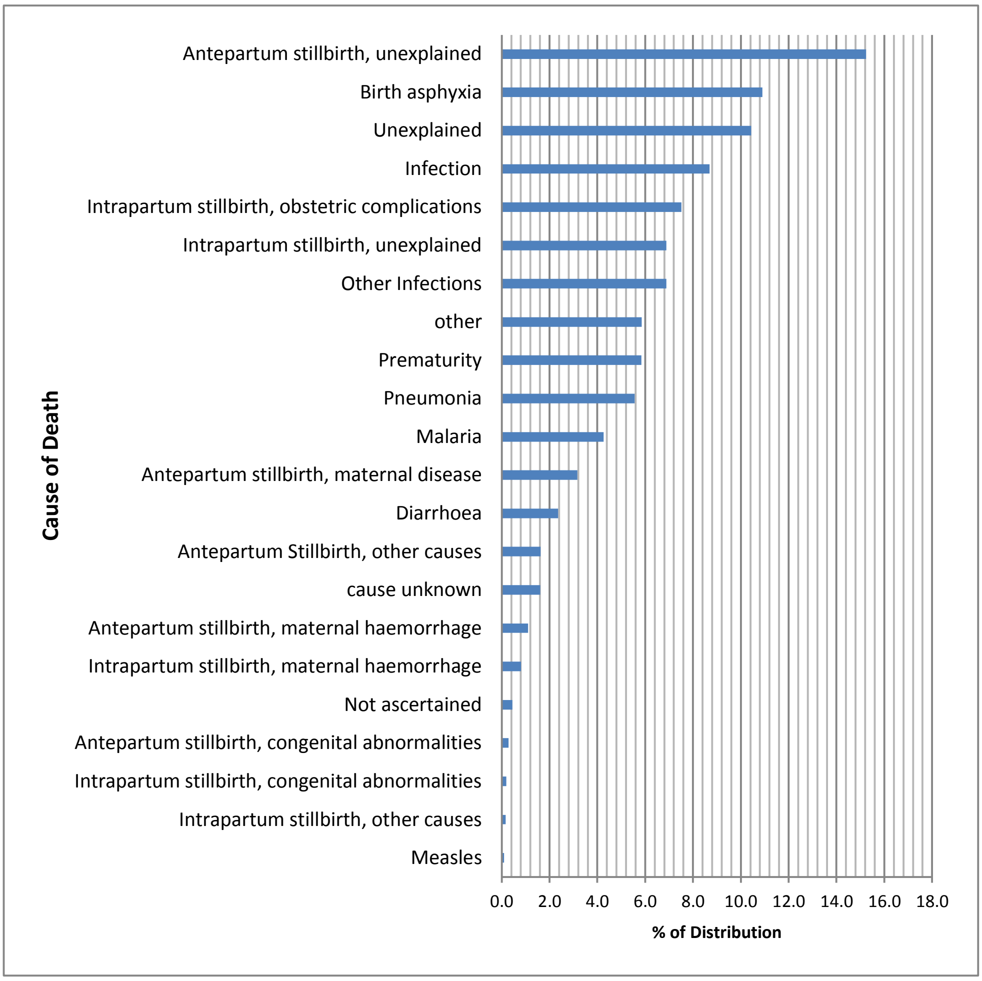


Figure 5.9 cause of death distribution among infants as annotated by physicians.

As seen from Figure 5.9 *Uncertain* is the most common category, encompassing over 20 per cent of the VAs that make up the women sub corpus. *Antepartum stillbirth, unexplained* is the leading cause of death in the infant sub-corpus, covering over 15 per cent. Note that the infant sub-corpus also has separate cause of death categories *Antepartum stillbirth, other causes*; *Unexplained; Other; cause unknown; Not ascertained*.

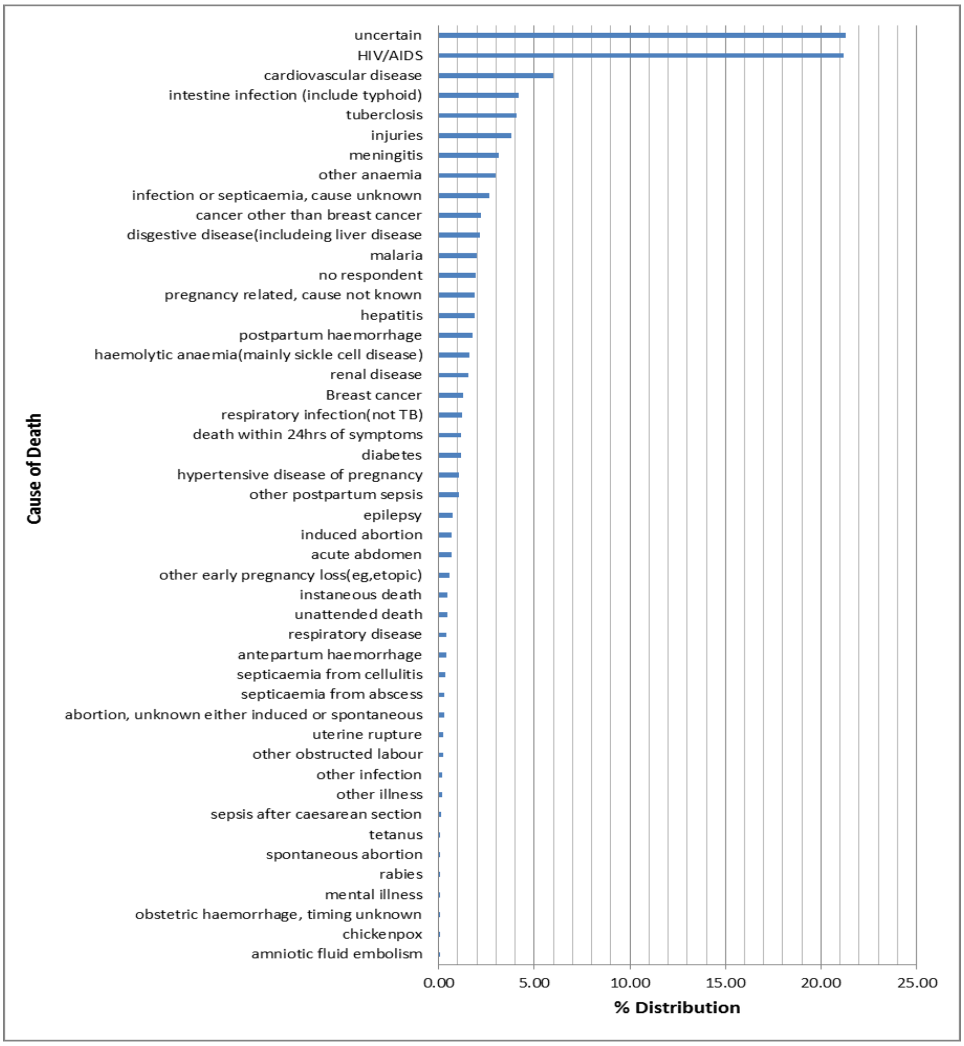


Figure 5.10 cause of death distribution among women as annotated by physicians

Figure 5.10 shows a rather different distribution of cause of death in the women sub-corpus. Two categories, *uncertain and HIV/AIDS* stand out as most significant, each being the cause of over 21% of deaths; then there is a long tail of other specific medical conditions, most of which are not found in the infants sub-corpus. The vague categories such as *Unexplained, other, cause unknown*, and *Not ascertained* do not feature in the women sub-corpus; we assume these are covered by the *uncertain* category. There are even rare cases of just one instance each such as *obstetric haemorrhage*, *timing unknown*; *chickenpox*; and *amniotic fluid embolism* found in the women sub-corpus. To a Computational Linguist without medical training, the groupings may not be clear with regards to how to distinguish between all categories and thus require special attention by language researchers when carrying out computational modelling.

This annotation scheme employed has a hierarchical structure, which is illustrated in Figure 5.11. The classification scheme takes into account reported medical symptoms and the time in which the death occurred.

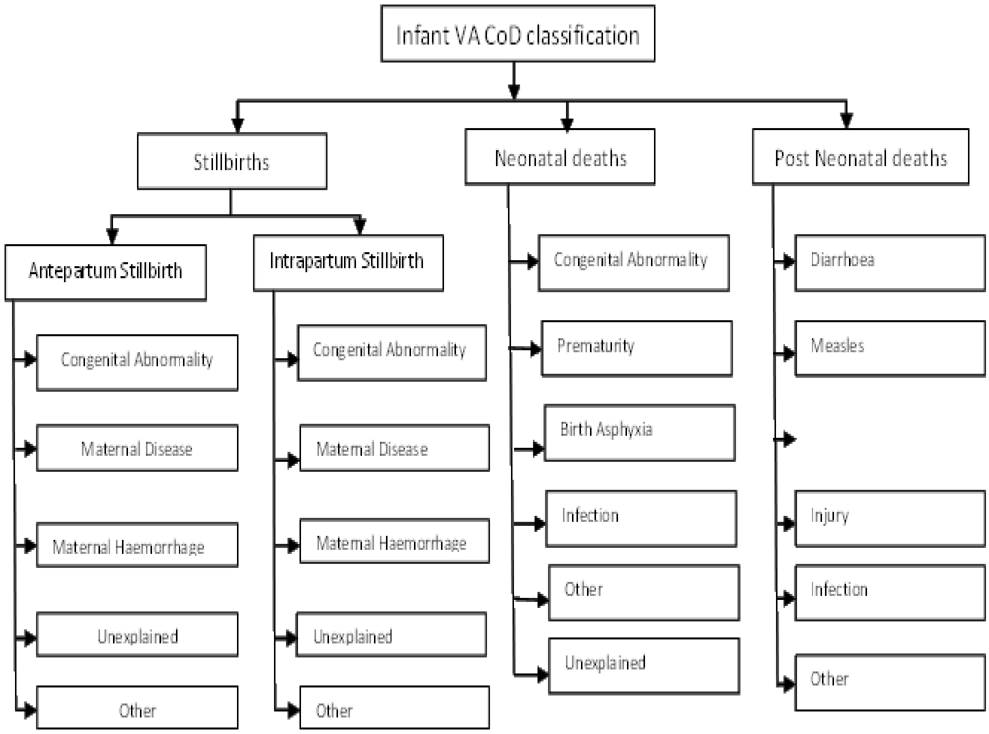


Figure 5.11 schematic diagram showing the hierarchy of causes of infant deaths adapted from Edmond et al. (2008)

As seen from Figure 5.11 *Congenital Abnormality* for example may be classified as either *Antepartum* or *Intra-partum* depending on whether the death occurred before the labour or during labour as explained by Edmond et al. (2008). Such groupings according to stage of delivery do not feature in the women sub-corpus cause of death annotations except for three causes of death involving haemorrhage at childbirth: *postpartum haemorrhage*, *antepartum haemorrhage*, *obstetric haemorrhage*, *timing unknown*.

5.5 Cause of Death Regrouping.

In Chapter Four it was discussed that Machine Learning algorithms tend to employ the Occam's Razor based principles, which were also demonstrated by Forman (2003) as showing that having a balanced text is a key requirement in Machine Learning tasks. This implies that a sample drawn from any domain of interest must give a balanced representation of that domain. For example, in language research, one might aim to give a balanced representation of text type: written versus spoken. Similarly, within Text Classification, an important issue to consider is the semantic categories or genre types. Within the context of VA, one of the semantic categories which required balanced representation was the cause of death. However, the original cause of death showed imbalanced distribution, which depicted the situation in the real world within the population where the data was obtained. We discussed the effect of imbalanced training data on the performance of a Machine Learning algorithm and some of the strategies found in the literature that are employed in mitigating those effects: oversampling and undersampling of training data, and cost sensitive learning. The former was referred to as a data level and the latter was also known as an algorithmic level strategy (Chawla et al., 2004). Considering the fact that some causes of death could not be assigned a definite cause of death label due to insufficient information available to the physicians during annotation, it is imperative to explore strategies to deal with these issues. We therefore propose a scheme to deal with these issues based on the principle of hierarchical groupings.

The principle of hierarchical groupings of category labels is not new to Computational Linguistics. This principle has been proposed as a means of dealing with ambiguity in word Part-of-Speech (PoS) categorisation by many language researchers, e.g. Knowles and Zuraidah Mohd (2003). Atwell et al.(1994) proposed this principle in Machine Learning research on unsupervised learning of word-clusters. It can be inferred from this that the deciding factor of any categorisation scheme for automatic systems is the usefulness of the scheme within the context of the intended application, which includes a cause of death classification scheme (Quigley, 2005) as discussed. This has resulted in various PoS-tagging schemes for English language, for example. The International Corpus of English (ICE) is an example of a corpus with multiple-level tag set (Greenbaum and Nelson, 1996); a PoS-tag consists of a broad category (e.g. noun) and a set of finer-grained subcategories (e.g. common singular noun). Drawing inspiration from the ICE tagging scheme, the semantic categories of the VA documents demonstrated in Figure 5.12 could have multiple categories: Time-of-Death and Type-of-Death. Figure 5.12 shows a proposed scheme to classify VA documents based on their content and the cause of death hierarchy.

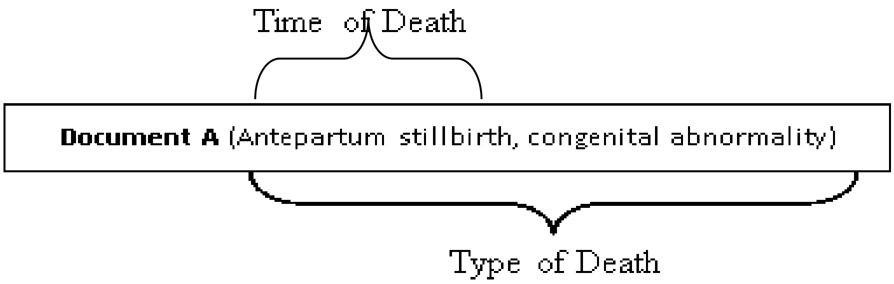


Figure 5.12 scheme to re-classify VA document based on hierarchy

The scheme as shown in this figure may enable documents to be classified at two levels: *Time-of-Death - Antepartum stillbirth or Type-of-Death - Antepartum stillbirth*, *congenital abnormality.* Having these levels of classification has an advantage for Machine Learning approaches to automatic classification. This is because the accuracy of Machine Learning models tends to increase with increase in number of training examples (Banko and Brill, 2001). In situations where there are rare cases of specific examples of cause of death in the training corpus, it may be reasonable to have a VA document classified at a higher level *(Time-of-Death*) than the fine grained level (*Type-of-Death)*.

Based on the above proposed scheme we re-grouped the *Original-groupings* into three different groupings: *Groupings1*, *Groupings2* and *Groupings3*. *Groupings1* was derived based on the hierarchy shown in Figure 5.11. *Groupings2 and Groupings3 were created* in reaction to consultation and advice received from Professor Betty Kirkwood and other public health experts from the London School of Hygiene and Tropical medicine. These groupings are discussed in turns.

5.5.1 Cause of Death Groupings1

Table 5.3a and 5.3b show the *Time-of-Death* and *Type-of-Death* categories for the proposed Groupings1 respectively. The Time-of-Death has five categories: *Intrapartum_stillbirth*, concerned with all deaths that occurred during delivery. *Antepartum_stillbirth* are deaths that occurred before delivery. *Non_stillbirth_unknown_cause* are all the deaths that could not be determined by the physicians. Neonatal are all deaths that occurred within 28 days of life and *PostNeonatal* are all deaths that occurred after 28 days. The *Type-of-Death* level has 16 categories:

Table 5.3a Proposed mapping of Time-of-Death categories for Groupings1

Original grouping	Time-of-Death
Intrapartum stillbirth, other causes	Intrapartum_stillbirth
Intrapartum stillbirth, congenital abnormalities	
Intrapartum stillbirth, unexplained	
Intrapartum stillbirth, maternal haemorrhage	
Intrapartum stillbirth, obstetric complications	
Antepartum Stillbirth, other causes	Antepartum_stillbirth
Antepartum stillbirth, congenital abnormalities	
Antepartum stillbirth, maternal haemorrhage	
Antepartum stillbirth, unexplained	
Antepartum stillbirth, maternal disease	
cause unknown	Non_stillbirth_unknown_cause
Unexplained	
Not ascertained	
Birth asphyxia	Neonatal
Infection	
Pneumonia	
Prematurity	
Other	
Other Infections	PostNeonatal
Measles	
Diarrhoea	
Malaria	

Table 5.3b Proposed mapping of Type-of-Death categories for Groupings1

Original grouping	Type-of-Death
Antepartum Stillbirth, other causes	Stillbirth_other_causes
Intrapartum stillbirth, other causes	
Antepartum stillbirth, congenital abnormalities	Stillbirth_congenital_abnormalities
Intrapartum stillbirth, congenital abnormalities	
Antepartum stillbirth, unexplained	Stillbirth_unexplain
Antepartum stillbirth, unexplained	
Intrapartum stillbirth, unexplained	
Antepartum stillbirth, maternal haemorrhage	Stillbirth_maternal_haemorrhage
Intrapartum stillbirth, maternal haemorrhage	
Intrapartum stillbirth, obstetric complications	Stillbirth_obstetric_complications
Antepartum stillbirth, maternal disease	Stillbirth_maternal_disease
Cause unknown	Non_stillbirth_unknown_cause
Unexplained	
Not ascertained	
Birth asphyxia	Birth_asphyxia
Infection	Neonatal_infection
Pneumonia	Pneumonia
Prematurity	Prematurity
other	Neonatal_other_causes
Other Infections	PostNeonatal_other_infections
Measles	Measles
Diarrhoea	Diarrhoea
Malaria	Malaria

We decided to keep the *Non_stillbirth_unknown_cause* category in this grouping due to the fact that negative examples tend to be useful to Machine Learning algorithms in order to discover the decision boundaries during learning for prediction (Hospedales et al., 2013). Thus, it was done under the hypothesis that keeping these causes of death together as one category might improve the algorithm's ability to distinguish between categories.

5.5.2 Cause of Death Groupings2

Table 5.4a and 5.4b show *Time-of-Death* and *Type-of-Death* level categories for *Groupings2* respectively.

Table 5.4a Proposed mapping of Time-of-Death categories for Groupings2

Original grouping	Time-of-Death
Intrapartum stillbirth, other causes	Intrapartum_stillbirth
Intrapartum stillbirth, congenital abnormalities	
Intrapartum stillbirth, unexplained	
Intrapartum stillbirth, maternal haemorrhage	
Intrapartum stillbirth, obstetric complications	
Antepartum Stillbirth, other causes	Antepartum_stillbirth
Antepartum stillbirth, congenital abnormalities	
Antepartum stillbirth, maternal haemorrhage	
Antepartum stillbirth, unexplained	
Antepartum stillbirth, maternal disease	
Birth asphyxia	Neonatal
Infection	
Pneumonia	
Prematurity	
Other	
Other Infections	PostNeonatal
Measles	
Diarrhoea	
Malaria	

Table 5.4b Proposed mapping of Type-of-Death categories for Groupings2

Original grouping	Type-of-Death
Intrapartum stillbirth, other causes	Stillbirth_other_causes
Intrapartum stillbirth, congenital abnormalities	Stillbirth_congenital_abnormalities
Antepartum stillbirth, unexplained	
Intrapartum stillbirth, unexplained	Stillbirth_unexplain
Antepartum stillbirth, maternal haemorrhage	
Intrapartum stillbirth, maternal haemorrhage	Stillbirth_maternal_haemorrhage
Intrapartum stillbirth, obstetric complications	Stillbirth_obstetric_complications
Antepartum Stillbirth, other causes	Stillbirth_other_causes
Antepartum stillbirth, congenital abnormalities	Stillbirth_congenital_abnormalities
Antepartum stillbirth, maternal haemorrhage	Stillbirth_maternal_haemorrhage
Antepartum stillbirth, maternal disease	Stillbirth_maternal_disease
Birth asphyxia	Birth_asphyxia
Infection	Neonatal_infection
Pneumonia	Pneumonia
Prematurity	Prematurity
other	Neonatal_other_causes
Other Infections	PostNeonatal_other_infections
Measles	Measles
Diarrhoea	Diarrhoea
Malaria	Malaria

Table 5.4a and 5.4b show the difference between *Groupings1* and *Groupings2* with the removal of the *Non_stillbirth_unknown_cause* category. As discussed in Chapter Four that noise from categories could potentially cause performance of a Machine Learning algorithm to deteriorate, and thus the removal was done to explore the effect of removing this category from the training set which could be a potential source of noise.

5.5.3 Cause of Death Groupings3

Table 5.5a and 5.5b show *Time-of-Death* and *Type-of-Death* level categories for *Groupings3* respectively. Time-of-Death consists of three categories: Stillbirth, Neonatal and PostNeonatal. The *Type-of-Death* moreover has a total of 10 categories.

Table 5.5a Proposed mapping of Time-of-Death categories for Groupings3

Original grouping	Time-of-Death
Intrapartum stillbirth, other causes	Stillbirth
Intrapartum stillbirth, congenital abnormalities	
Intrapartum stillbirth, unexplained	
Intrapartum stillbirth, maternal haemorrhage	
Intrapartum stillbirth, obstetric complications	
Antepartum Stillbirth, other causes	
Antepartum stillbirth, congenital abnormalities	
Antepartum stillbirth, maternal haemorrhage	
Antepartum stillbirth, unexplained	
Antepartum stillbirth, maternal disease	
Birth asphyxia	Neonatal
Infection	
Pneumonia	
Prematurity	
Other	
Other Infections	PostNeonatal
Measles	
Diarrhoea	
Malaria	

Table 5.5b Proposed mapping of Type-of-Death categories for Groupings3

Original grouping	**Type-of-Death**
Intrapartum stillbirth, other causes	Intrapartum_stillbirth
Intrapartum stillbirth, congenital abnormalities	
Intrapartum stillbirth, unexplained	
Intrapartum stillbirth, maternal haemorrhage	
Intrapartum stillbirth, obstetric complications	
Intrapartum stillbirth, obstetric complications	
Antepartum stillbirth, maternal haemorrhage	Antepartum_stillbirth
Antepartum stillbirth, unexplained	
Antepartum Stillbirth, other causes	
Antepartum stillbirth, congenital abnormalities	
Antepartum stillbirth, maternal haemorrhage	
Antepartum stillbirth, unexplained	
Antepartum stillbirth, maternal disease	
Birth asphyxia	Birth_asphyxia
Infection	Neonatal_infection
Pneumonia	Pneumonia
Prematurity	Prematurity
other	Neonatal_other_causes
Other Infections	PostNeonatal_other_infections
Measles	
Diarrhoea	Diarrhoea
Malaria	Malaria

As the Tables 5.5a and 5.5b show, all *Stillbirth* related deaths were grouped together for the *Time-of-Death* but were split into *Intrapartum_stillbirth* and *Antepartum_stillbirth* for *Type-of-Death*. It is also notable *Measles* and *Other_infections* have been combined to become *PostNeonatal_other_infections* as one of the *Type-of-Death* categories since they are all infection.

5.6 Summary

This Chapter described a system employed to effectively manage the surveillance system which covered an area of over 240,000Km^2 of seven districts in rural Ghana. This facilitated the process of identifying deaths of adult women and children within a population of over 200,000. The methods employed as part of the data management strategies to effectively and efficiently manage the inherent complexities of the surveillance system were described.

Furthermore, we described in this Chapter the methods employed in building an annotated corpus which is suitable for the development of automatic methods for the analysis of VA and other language research. The methods employed in collecting and building the corpus have also been presented and include quality measures taken during the annotation process to ensure good quality annotation was obtained.

Finally, we analysed the cause of death annotations obtained from the corpus and proposed three different classification schemes based on a set of principles and expert advice which will subsequently form the basis of our experiments.

Chapter 6

Analysis of the Verbal Autopsy Corpus

6.1 Introduction

McEnery and Wilson (1996) described a modern corpus as a collection of text which has several characteristics: machine readability; representative samples; a finite size; and being a standard of reference for the language of the domain of which it represents. The attributes are expanded further to cover the storage format of the annotated corpus; and to define annotation levels. These attributes have been found to have effect on the use of corpora beyond their original purpose of creation (Cohen et al., 2005).

In Chapter Five we described the corpus building process and we demonstrated that the corpus had been encoded in a format readable by machines as required. In this Chapter we analyse the corpus to demonstrate that the VA corpus has the general characteristics of a modern corpus as described above.

6.2 Choice of format and encoding standards

There are several choices available as to the format and standards in which a corpus can be processed and stored. However, the overriding principle should be to allow users the flexibility and the freedom to manipulate and access annotation information. A survey conducted by Cohen et al. (2005) on corpora formats suggested that none of the non-XML formats met recoverability criteria. In other words, a corpus encoded in any format other than XML can pose difficulties mapping between the annotations and the original text. XML encoding was also the recommended strategy as it enabled standoff storage, where the annotations were stored separately from the original text (Leech, 1993). The advent of various software tools and standard libraries in the application programming interfaces of standard programming languages such as Java and Python has made both inline and stand-off annotation with XML a more convenient encoding approach. This justified the decision to encode the VA corpus in an XML format with inline annotation where both annotation and original text were stored in a single file as shown in Figure 5.8. We opted for the inline strategy as it was much a convenient for our task.

6.3 Difference between Infant and Women sub-corpora

Figures 6.1 and 6.2 show a wide spread of the length of the documents which make up the corpus.

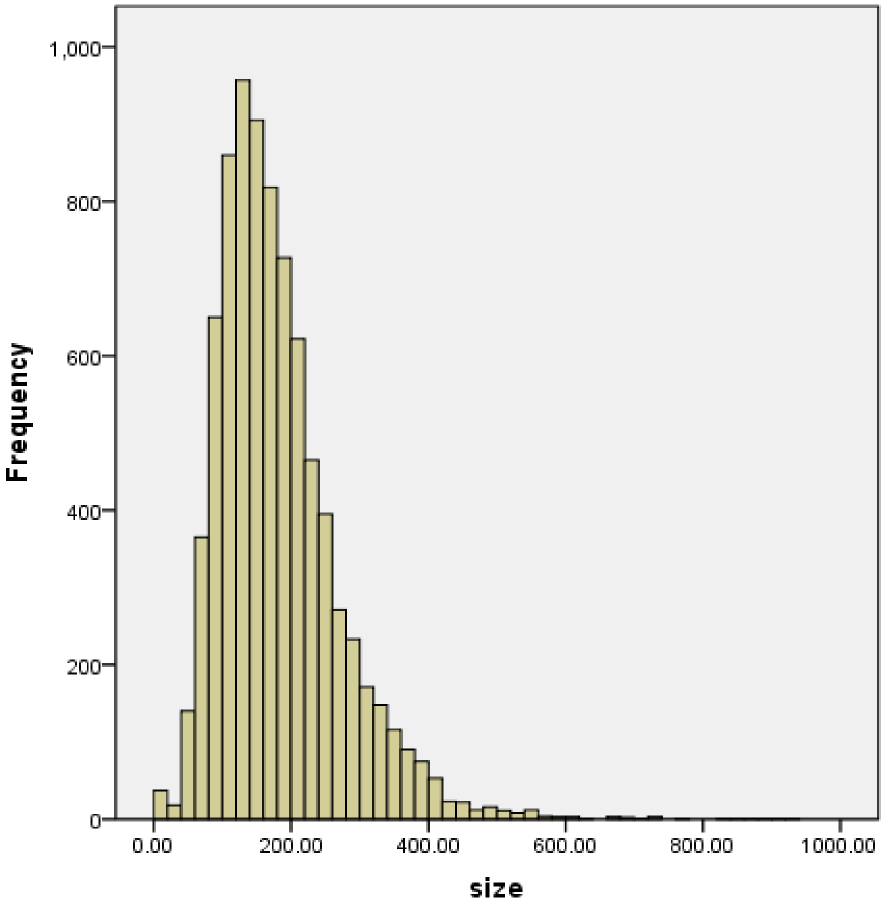


Figure 6.1 Distribution of document size of the infant sub corpus

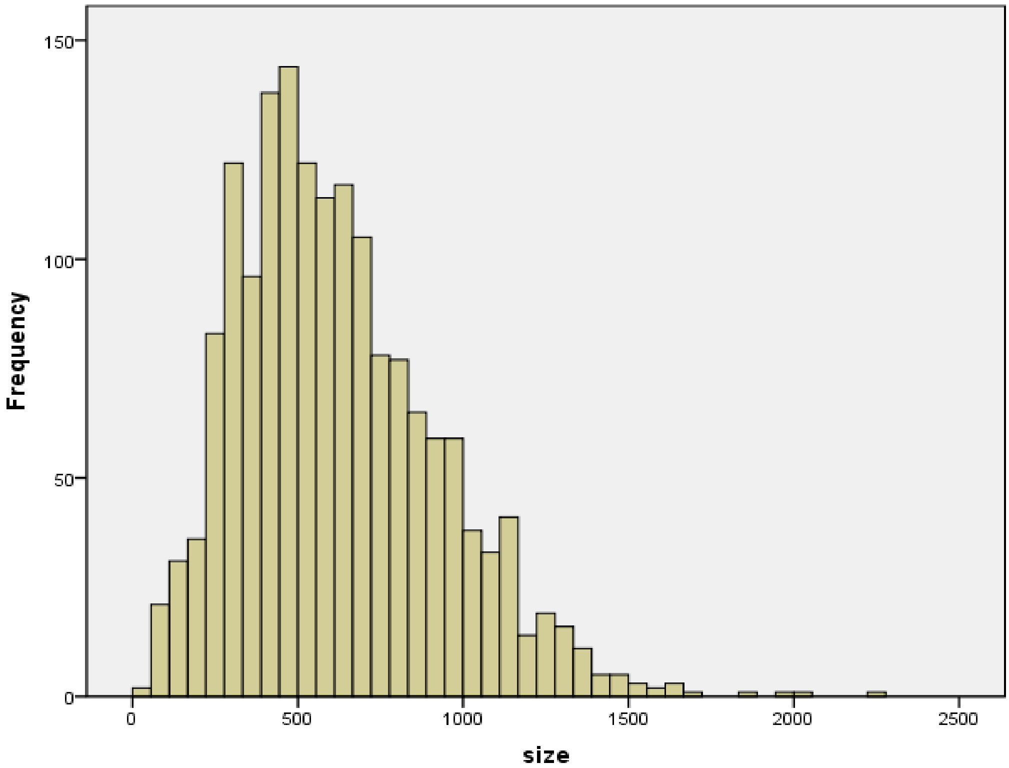


Figure 6.2 distribution of document size of the women sub corpus

Figure 6.1 indicates a spread of the infant sub-corpus, ranging between 15 and 550 words, with an average of 182. Similarly, Figure 6.2 also gives an indication of the spread of document length, ranging between 33 and 2200 words, with an average of 630 words. As already pointed out that the women questionnaire contained more questions than the infant one, and so it is not surprising to see from the Figures that the women document length was greater. Variation in document length plays a key role in computational methods and this characteristic is worth noting. For example, in document classification, it is important to consider normalisation during feature representation, which takes into account the length of the document (Leopold and Kindermann, 2002).

The infant and women log-likelihood comparison: Table 6.1 shows a log-likelihood key-word comparison between the infant and women sub corpora.

Table 6.1 Log-likelihood key-word comparison between Infant and Women sub corpora

Infant		Adult	
Log-likelihood	**Word**	**Word**	**Log-likelihood**
18457.747	i	she	19381.94
5003.486	baby	her	4813.075
2824.311	my	deceased	3800.457
2426.954	child	complaining	3110.407
1946.707	birth	complained	2451.803
1939.428	labour	hospital	1264.405
1554.711	delivered	admitted	1198.602
1508.423	pregnancy	sent	1109.615
1475.825	delivery	later	932.191
1288.805	he	years	926.846
1173.189	me	for	883.649
906.209	during	died	880.077
849.069	any	complain	737.514
779.007	when	became	728.608
738.985	normal	admission	677.243
727.43	born	heart	404.027
628.19	him	typhoid	267.469
589.087	did	hypertension	120.225
548.812	immediately	accident	118.613
538.709	average	cancer	110.498

The table further demonstrates some of the similarities and differences that exist between the two sub corpora. As observed from the table, there were similarities in the use of pronouns, which tended to feature prominently in both corpora. For example, *I* was the most frequent word in the infant sub-corpus whereas *she* appeared as the most frequent word in the women sub-corpus. This suggests a true reflection of what pertains in the discourse during VA interview. The *'I'* appearing as the most frequent word in the infant sub corpus also suggests that the mother of the child was the one who mostly gave account of what led to the death of the child. This narrative began with the events that surrounded pregnancy. In contrast, the most frequent word in the women sub corpus *she* refers to the deceased, suggesting the relative giving the account was not as directly involved in events leading to the death. Observing the top 100 words from both corpora further differences can be established. Notably *labour*, *delivery*, and baby are all words that described the infant corpus, while in contrast words such as complained, accident, hypertension and cancer were words that frequently occurred in the women sub corpus.

6.4 The Verbal Autopsy corpus and Zipf's law

The VA corpus word frequency distribution was assessed against Zipf's law, also known as a Power-law distribution, which was found to occur in a range of phenomena including words in human language (Newman, 2005). Zipf's law establishes the relationship between the frequencies of words and the rank in which they occur in the list (Manning et al., 1999), where rank number 1 represents the most frequent word in the corpus. This can be expressed mathematically as: R * F ≈ C, where R is the rank of a given word, and F is the frequency of that word and C is an approximately constant value which depends on the size the corpus being examined. Using Python scripts, all the text was extracted from the XML file into a plain text file. A natural language analysis toolkit called AntConc (Anthony, 2005) was applied to the entire plain text file to generate word frequencies and their ranks. The output of this was used to plot a graph in Figure 6.3, which shows Zipfian word distribution in the VA corpus transformed on a logarithm scale.

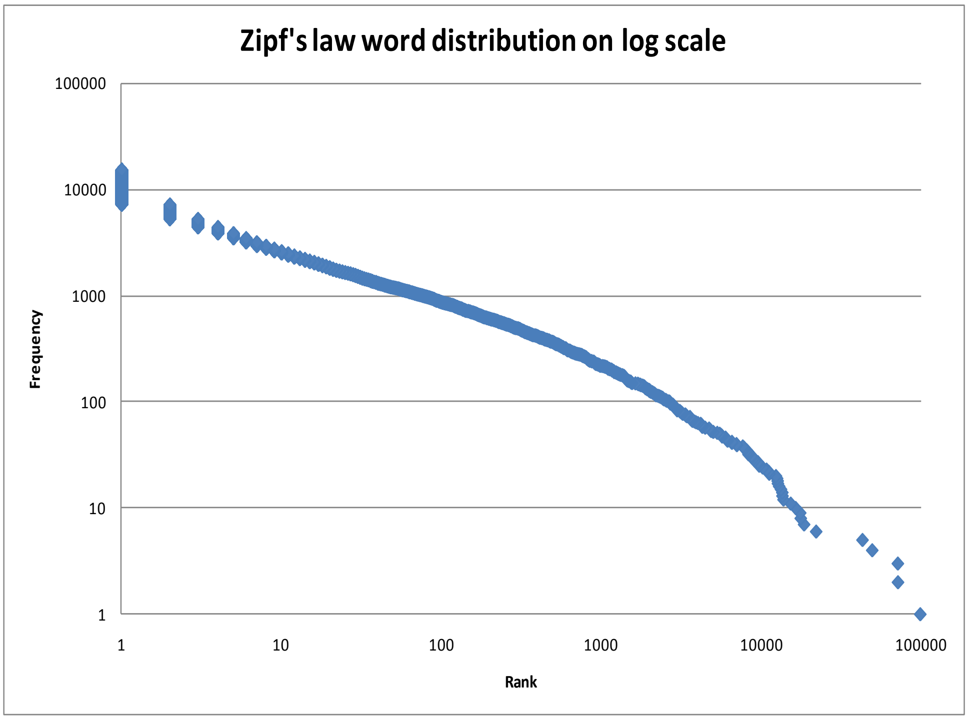


Figure 6.1 word frequency distribution of the VA corpus on log scale

As can be seen the graph has a negative slope, which suggests the word distribution phenomenon expected from Zipf's law. The proportion of words with higher ranks but very low frequency has implications for lexical modelling, as one would have to deal with this high level of lexical sparseness. This characteristic is tested later in this Chapter.

6.5 Sparseness and Lexical Diversity

Another property that is exhibited in most natural language text, as a side-effect of Zipf's law is the issue of sparseness (Goweder and De Roeck, 2001). A corpus is said to be sparse if many of the words in the corpus are uncommon or unknown. It can also mean that some word-combinations or ngrams are rare in the corpus (Jurafsky and Martin, 2008). The figures shown above suggest a high level of sparseness and imbalance in the vocabulary and also in the cause of death categories, which could be problematic when trying to apply Machine Learning: classifiers work best when features of the data and classification categories are reasonably frequent and balanced (Lakeland and Knott, 2004).

An assessment was carried out using the word type-token ratio to ascertain the lexical variability that existed in the VA corpus (Youmans, 1991). The word type-token ratio is defined as the number of distinct words in a given text divided by the length of the text, and should have a value between one and zero. A ratio approaching one is a high ratio, which means that each word occurs only once in the text, and therefore the text is lexically sparse. Conversely, if the type-token ratio is a small fraction close to 0, then the text has low sparseness, which implies that there are a few words repeated many times in the text, which should result in a closed vocabulary text. To test this, we applied the experiment originally carried out by Yahya (1989) to observe the behaviour of word occurrence patterns in Arabic with respect to English, further adopted by Goweder and De Roeck (2001). For this experiment, we compared the type-token ratio of samples from our corpus with figures based on equivalent-sized samples from the Brown corpus (Francis and Kucera, 1979). Some argue that the type-token ratio experiment should exclude stop-words (Graesser et al., 2004). However, as pointed out by Forman (2003), stop-words could be domain specific since what is considered a stop-word in one domain could be relevant in another domain. For example the word *during*, which is regarded as a function word but also forms part of an expression *during labour* could be the discriminative word that could predict a document belonging to *Intra-partum* as opposed to *Ante-partum* category. The experiment was therefore carried out with all the words.

Table 6.2 Type token ratio derived from the infant sub-corpus

Length of text	VA corpus distinct words	VA corpus word ratio	Brown corpus distinct words	Brown corpus word ratio
100	63	0.63	69	0.69
200	109	0.54	124	0.62
400	177	0.44	165	0.41
800	279	0.38	328	0.41
1,600	398	0.24	621	0.38
3,200	596	0.18	871	0.27
6,400	839	0.13	1,361	0.21

In Table 6.2 the word type-token ratio revealed an interesting characteristic of the VA corpus. It can be observed from the table that the majority of the values for the ratios were lower compared to the corresponding values found in the Brown corpus, which supports the intuition that the VA corpus has a relatively closed vocabulary since the Brown corpus comprised of a variety of genres and more general text. This view seems to hold as the sample text length increases. For example the text length of 6,400 had ratios of 0.13 and 0.21 for VA and the Brown corpus respectively, which suggests that the Brown corpus tends to be sparser or contains more uncommon words than the VA corpus. However, it could also be observed that there is still a significant amount of vocabulary variability in the VA corpus samples due to a variety of reason such as spelling issues; use of nonstandard and variety of ways of expressing concepts as outlined in table 6.3.

6.6 Linguistic Complexity of the Verbal Autopsy Language

The quality of the text introduces challenges for automated Text Analyses. Table 6.3 is a summary of the issues identified in VA text with instances of spelling errors and non-standard forms of expressing medical concepts. For example the expression "“*I visited xxx hospital on Tuesday and was given one bottle of water*.” is describing a saline drip that was given to the child when the child was sent to the health facility for treatment. These expressions are a potential source of noise and data sparseness. The consequence of these results are varying degree of noise and variance, which has a further adverse impact on the performance of Machine Learning algorithms as discussed in Chapter Four.

Table 6.3 Catalogue of types of issues in Verbal Autopsy open narrative text

Type of issue	Example
Grammar and spelling errors.	"*Before labour waters, which look clear and without bad scent"* *"… she fell sick, which lauted for three days*.."
Colloquial forms in expressing concepts	Baby came out Baby landed Gave birth Delivery
Use of local terms to describe medical conditions	Asram, Anidane
Non-standard expressions of medical concepts	"*I visited xxx hospital on Tuesday and was given one bottle of water*. .."
Abbreviations and acronyms	*TBA* = Traditional Birth Attendant *ANC* = Antenatal care .
Inappropriate use of punctuation marks	"*Any time, she breaths, you see a hole*"

6.6.1 Performance of Part of Speech Taggers on VA Text

PoS tagging is a standard annotation added to corpora, and the accuracy of the PoS tagging is important to achieving reliable analyses. However, accuracy of PoS tagging is dependent on the quality of the text. We examined this issue by assessing the accuracy of standard PoS-taggers, from the Natural Language Tool Kit - NLTK (Loper and Bird, 2002). To achieve this, three PoS tagging algorithms were trained with a standard PoS-tagged English corpus, the Brown corpus (Francis and Kucera, 1979), available as part of the NLTK. These were subsequently used to tag a 2000 words sample randomly generated from the corpus.

The three PoS tagging algorithms employed were: Hidden Markov Model-HMM (Huang, 2009); Trigram and Tags - TnT tagger (Brants, 2000); and Brill Transformational-Based Learning algorithm (Brill, 1992). Even though these are considered state-of-the-art and widely used, their underlying assumptions and approaches to tagging differ. The HMM and TnT are stochastic-based algorithms which tend to use similar approaches to tagging, as both analyse the sequential history of word–tag pairings in a given ‘sentence’ using Markov Model principles (Ghahramani, 2001). However, the TnT tagging approach differs from HMM based on the features it employs. For example, unlike HMM, orthographical features such as capitalisation are taken into account by the TnT algorithm. Moreover, the Brill tagger, is an example of Transformational-Based Learning (TBL) algorithm. Like a stochastic tagger, it begins by pairing words with their most frequent tag observed from the corpus. It then applies rules derived from the corpus to correct the output of the stochastic tagger. Due to this process requirement, a unigram algorithm was also employed as a base.

The outputs of the taggers were inspected by a PoS-tagging expert (Atwell), see Atwell et al. (1994) and incorrectly tagged words were marked. Percentage accuracies of the taggers were then calculated as percentage of the number of correctly tagged words out of the total number (2000).

The output of each of the three PoS taggers was very similar with no significant difference in the results observed. The average performance accuracy obtained for each of the PoS-taggers on the VA corpus was 83 per cent. This result was much lower and a significant departure from the performance of PoS taggers on general “well-formed” English text, which is generally reported as about 96–97 per cent (Brants, 2000).

6.7 The effect of annotation

The VA corpus lacks linguistic or syntactic annotations; the Cause of Death classification of each VA is a medical semantic annotation. The findings of a survey suggested that a significant number of biomedical corpora tends to have no linguistic annotations (Cohen et al., 2005). This does not rule out the need for linguistically annotated corpora since this information could increase their usefulness, as evidenced by the popularity of some corpora over others within the biomedical domain. The GENIA corpus (Kim et al., 2003) for example has been employed in numerous tasks outside its original purpose in which it was built. GENIA has linguistic annotations in addition to semantic annotations.

Given that our VA corpus lacked other linguistic and syntactic level annotations, we carried out an experiment to identify potential keywords that best described our semantic annotations. The experiment considered our proposed *Groupings1* discussed in Chapter Five. This experiment in other words could be described as feature extraction, which is well-established technique in computational modelling (Worzel et al., 2007). The underlying hypothesis of the experiment was that there were words which should be specific to a particular cause of death category. These words could be the symptoms mentioned in the text. This experiment was meant to illustrate the feasibility of using corpus based methods to analyse and identify keywords in a text that could correlate with a given cause of death category. The experiment focused on the most common cause of death categories of infant VAs. The method employed in this experiment is fully described in Chapter Seven of the thesis. Table 6.4 shows a sample output obtained from the method.

Table 6.4 Log-likelihood results obtained from Time-of-Death categories for Groupings1

Intra-partum Stillbirth		**Antepartum Stillbirth**		**Neonatal**		**PostNeonatal**		**Non_Stillbirth Unknown cause**	
Word	LLH	Word	LLH	word	LLH	Word	LLH	Word	LLH
dead	410.6	dead	485.0	weak	240.9	he	138.3	bed	77.4
still	277.9	i	386.5	cry	195.6	healthy	107.6	him	58.5
i	246.6	womb	269.9	breathing	188.3	sent	95.0	adwoa	57.3
out	143.7	movement	205.9	could	121.5	diarrhoea	86.2	healthy	55.6
coming	95.0	stillbirth	191.9	died	119.5	bought	73.3	any	54.7
stillbirth	91.9	macerated	189.6	breath	110.6	hot	73.3	problem	36.2
came	80.4	still	176.7	incubator	108.3	him	71.7	kandege	34.6
already	77.9	already	153.2	machine	103.9	coughing	71.6	wake	34.0
deliver	63.0	my	109.3	oxygen	99.3	her	66.5	slept	33.5
operation	55.5	moving	108.7	minutes	88.4	she	63.7	without	33.5

Table 6.4 shows the top ten words obtained from the experiment based on *Groupings1 Time-of-Death* categories. As seen from the table, the experimental result supported the hypothesis. All the top ranked words could intuitively be associated to the cause of death categories. For example it was not surprising to see that *dead*, *already*, *coming* and *stillbirth* appeared as the top ranked words for *Intra-partum stillbirth* but did not appear in the *Non Stillbirth Unknown Cause* category. Dead could be an indication of a baby not born alive and coming could be an indication of Intra-partum, suggesting that the baby died during delivery. Similarly, words such as cry, breathing, incubator, and oxygen were strong indicators for *Neonatal* death, suggesting the baby was born alive because it cried, but died later after it was kept in an incubator. These words however, did not appear in the *Intra-partum stillbirth* or *Antepartum stillbirth* categories.

6.8 Discussion

We have also demonstrated from our analysis the feasibility of employing corpus based approaches in the analysis of this corpus by identifying keywords associated with a given cause of death category. The experiment focused on the infant sub-corpus and on the high level of cause of death grouping.

The comparison of the lexical diversity between the VA and the Brown corpora was based on convenience since the type-token ratio for the Brown corpus was already available to compare with. This may not be the most appropriate comparison considering the content of the Brown corpus is general English text. It could be interesting to compare these results with a corpus from the biomedical domain. That however does not invalidate the findings about the lexical diversity of the VA corpus using the type-token ratio between zero and one as the reference range.

Several factors could account for the low performance of the PoS taggers. One possible factor was the language used in the VA reports. As table 6.2 demonstrated, the grammatical structures found in the VA text differ from those found in standard English text such as news. For example, "*before labour waters, which look clear and without bad scent broke*" was a typical sentence in the VA narratives. This example clearly showed the grammatical problems present in the text required to be handled by the PoS tagger. Other factors included: wrong use of punctuation; and unknown words. For example *intravenous* was not found in the training corpus and therefore resulted in a wrong PoS tag of plural noun being assigned by the PoS tagger as the default tag for words ending with -s. This further demonstrated the complexity of the language used in VA corpus. This raises questions about the robustness of PoS taggers developed and evaluated using standard English corpora such as Brown, LOB (Johansson et al., 1986), or Wall Street Journal (Paul and Baker, 1992) corpora. The performance suggested the need for a re-training of the state-of-the-art PoS algorithms with more appropriate corpora, such as the GENIA corpus, which contains a mixture of general news corpus plus biomedical domain corpus to assess their ability to deal with a wider range of text with various degrees of language challenges such as the VA text. Our experiment was conducted using the Brown corpus to train the PoS tagger, and potentially accuracy could have been improved if we had obtained a VA corpus annotated with PoS tags to serve as a training set for the PoS tagging process. This may be a useful approach for the re-evaluation and adaption of PoS taggers for VAs.

The high level of imbalance found in the cause of death distribution might be partly due to the methods employed in collecting the samples that form the content of this corpus. The approach adopted was natural as this process replicated a real word scenario. The rare cases could be a reflection of what pertained in the population in which the data was collected. One option, based on background information, might be to target populations with high incidence rates of the rarer cases in order to maximise the chances of identifying these rare categories to make up for the numbers towards achieving a balanced corpus. This could be part of the preparatory phase of the corpus process collection process.

The hierarchical classification scheme proposed may be described as a radical way of dealing with the skewedness found in the corpus. However, this approach may also be helpful to potential users once a clue can be given regarding the category of a specific VA document. For example having information that a child was born alive but died later facilitates the investigation process when compared with having no information at all.

Considering the fact that over 40 countries employ VA as an alternative approach to determining cause of death, this corpus creates a new dimension to language research within the biomedical domain. This new area of research could be useful to these countries. A future work could be to draw experiences from this project to build a corpus of multilingual content covering a range of countries that use VAs.

6.9 Summary

In this Chapter, an analysis of the VA corpus was conducted using various formal methods to describe and evaluate the corpus suitability for language research. The characteristics associated with the corpus were also described and their potential impacts on Machine Learning experiments were discussed.

Chapter 7

An exploration of the problem space of the Verbal Autopsy open narratives text

7.1 Introduction

This Chapter explores the open narrative of the VA text. We start by establishing a baseline, which allows performance obtained from further experiments to be computed and contextualised. We also carry out experiments on various aspects of the Text Classification processes in order to identify suitable methods for classification of VA text. The experiments involved the evaluation of various feature value representation schemes and Machine Learning algorithms. Unlike the baseline simple majority experiment which considers only the distribution of the categories in the corpus, the experiments to identify the best performing methods mentioned considered the unigram feature, which is a set of unique words found in our corpus.

Additionally, in Chapter Four we discussed issues with datasets that result in high dimensionality within the context of Machine Learning. Furthermore, in Chapter Six we demonstrated the possibility of high dimensionality associated with the VA corpus as a result of a variety of issues associated with the text. In this Chapter we propose a method named *locally-semi-automatic* approach as a strategy to address these problems. We examined the performance of this approach compared to some of the existing approaches. The experiments carried out in this Chapter are based on *Groupings1* cause of death categories.

It must be noted that the experiments described in this Chapter and the remaining thesis are based on only the infant sub-corpus. The rationale for this is that since the infant sub-corpus had a bigger sample as demonstrated in table 5.1 and also relatively smaller number of cause of death categories compared to the women sub-corpus, it was prudent to carry out the experiments on the infant sub-corpus as a proof of concept. Therefore, the hypothesis is that our approach should be generalisable when applied to the women sub-corpus. However, this is not tested in this thesis due to time constraint and recommended as a future work.

7.2 Pre-processing

During pre-processing, the text was converted to lower case and tokenised by whitespaces. All punctuation was also removed. Even though stop-words are removed during the pre-processing stage in most NLP tasks under the pretext that they are not informative and subsequently non discriminative, this removal still has led to mixed and inconclusive results in the past (Riloff, 1995). Also, Forman (2003) argues that stop-words tend to be domain specific, so the stop-words were therefore not removed from the dataset prepared from the corpus for this experiment.

7.3 Baseline

We employed a simple majority algorithm and applied it to the various *Groupings* of *Time-of-Death* and *Type-of-Death* categories as baseline algorithm. This algorithm predicts only the majority category of a given dataset (Witten and Frank, 2005). This was important in determining the best performing algorithm in relation to the simple majority algorithm. Thus, the expectation is that the selected algorithms and the models to be developed are expected to outperform the baseline algorithm.

7.4 Feature Value Representation Schemes and Algorithm Selection

In Chapter Four, we discussed some selected Machine Learning algorithms employed in classification tasks. These are: *Naive Bayes, Random Forest* and *Support Vector Machine* (SVM). The selection of these algorithms was justified as having been previously successfully applied to similar tasks and domain: The underlying principles of *Naïve Bayes* were employed by Byass et al (2006) to develop the *InterVA* as an automatic approach to analysis of VA based on the closed response part. Flaxman et al (2011) also employed *Random Forest* to develop an automatic approach to analysis of the closed response part of VA. SVM has also been successfully used in similar tasks and in domains in general (Cohen, 2006). However, this is the first research to apply SVM to the VA domain. The objective of this experiment was to investigate which of these algorithms would be suitable for classifying cause of death from VA open narrative text. Additionally, the feature value representation scheme investigations carried out in this thesis considered the standard term weighting schemes: *Binary*; *Term Frequency*; *TFiDF*; and *Normalised Term Frequency*, which are normalised by the length of the VA document due to the varying length of the VA documents (Danso et al., 2013a).

7.5 Curse of Dimensionality: The Cure

This section describes methods employed to solve the problem of the high dimensionality observed in the corpus as already discussed in Chapter Six. We implemented methods referred to as *locally-semi-automatic* and compared it with two standard approaches: *String matching and Information Gain*. The output of this experiment forms the basis for subsequent experiments described in this thesis. Since the objective of this experiment was to determine the most discriminative terms among the word unigrams, the output of this procedure will be referred to as *Discriminative Word Units* (*DWU*).

7.5.1 Our Locally-Semi-Automatic Approach

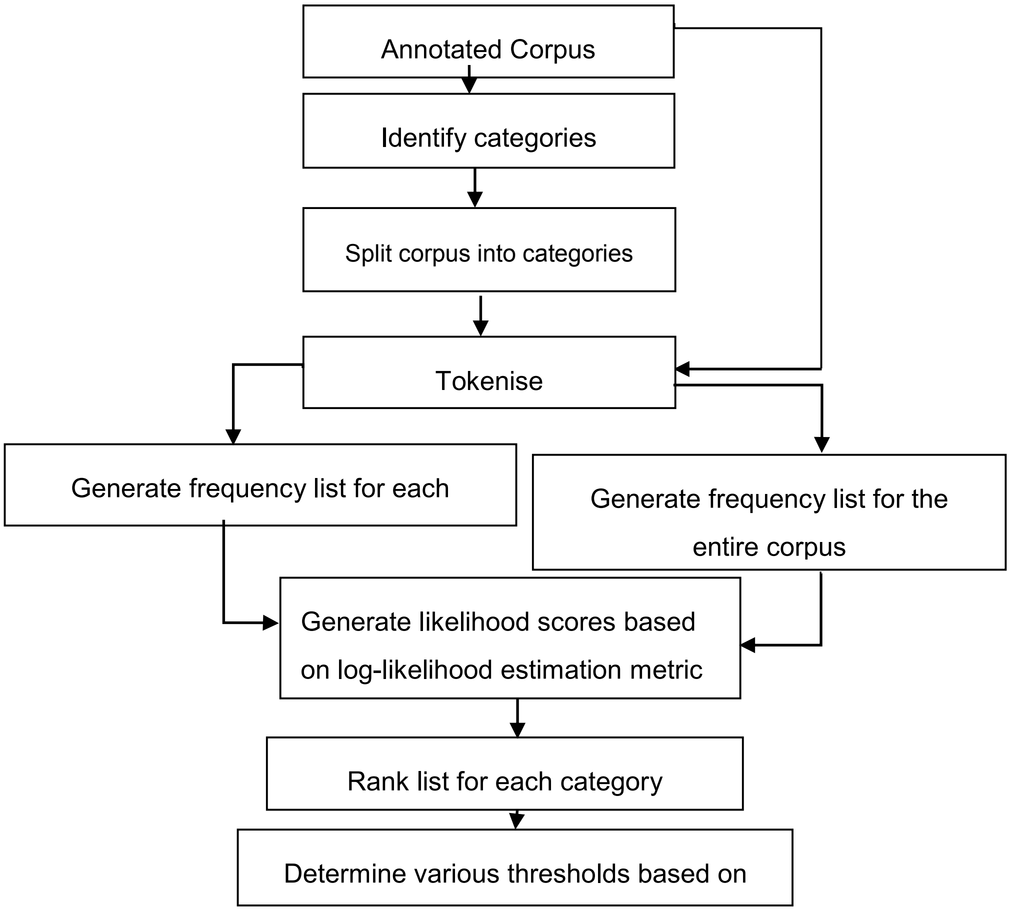


Figure 7.1 the locally-semi-automatic approach to feature reduction

The figure shows the procedure of the proposed method as described below:

I. As seen from the figure, the method requires a corpus annotated with some categories of interest.

II. The second stage is to identify these categories.

III. The next stage is to divide the corpus into separate files according to the identified categories.

IV. Each of the sub-corpora and the entire corpus are separately tokenised by whitespaces.

V. A frequency of occurrence list is generated for each tokenised file. A log-likelihood score is calculated for each sub-corpus (category) using the entire corpus frequency list as a reference corpus to determine the key-words associated with each category. This must be repeated in turn for the number of categories.

VI. The next step is to rank the likelihood list generated for each category by the likelihood score as shown in table 6.3 in Chapter Six.

To facilitate the determination of the keywords for each category, we employed the log-likelihood estimation metrics (Rayson and Garside, 2000) due to its superiority to other metrics as pointed out by Dunning (1993), which has been implemented as part of the AntConc software (Anthony, 2004). Various thresholds (top 10, 25, 50, 100, 150, 200, 250, 300, 350, and all words) were selected based on the rankings generated for each category and combined for the experiment.

7.5.2 String Matching Using Medical Lexicon

This approach explored the possibility of extracting medical terms mentioned in the VA open narrative text. This was based on an assumption that the medical terms used in describing the symptoms of deaths should be enough for a Machine Learning algorithm to learn and predict from. Reading samples of VA open narrative text suggested the medical terms used in describing the symptoms were simple and non-technical compared to the terms used in the clinical or biomedical settings. The medical terms found in the corpus could be described as medical terms used in everyday language - similar to those found in an online "kids" medical dictionary. From the findings a list of the terms was compiled from the online "kids" medical dictionary into a lexicon of lay medical terms. The compiled lexicon contained 522 unique medical terms. A string matching look-up algorithm was employed to extract mentions in the VA text used in describing cause of death. Only fully-matched terms were retrieved, and these were employed as features for classification.

7.5.3 Feature Selection by Information Gain

Information-gain-based approach is a global supervised feature selection method that can be used to select features based on the Information Gained metric as described in Forman (2003). This algorithm has an implementation available in WEKA (Witten and Frank, 2005) and remains one of the standard and popular feature reduction methods employed as a pre-processing step for the classification task (Witten and Frank, 2005). We employed this in our experiments in order to determine how it performs against our proposed method.

7.6 Experimental Setup

Separate datasets of unigrams were prepared based on the feature value representations under investigation: *Binary*, *Term Frequency*; *Normalised Term Frequency*, expressed as the *Term Frequency* divided by the total number of terms found in the given document (document length); and TFiDF. The files were converted and stored into .arrf[1], which is a format readable by the WEKA used in carrying out this experiment. WEKA has implementations of the Machine Learning algorithms discussed in Chapter Four, and these algorithms were employed in carrying out the experiments: the Naïve Bayes algorithm developed by John and Langley (1995); the Platt's Sequential Minimal Optimisation(SMO), which is a variant of the standard SVM algorithm (Keerthi et al., 2001); and the Random Forest, also a variant of the Standard Decision Tree algorithm (Breiman, 2001). Furthermore, the feature reduction strategies described above were applied to the best performing algorithm based on the unigrams.

7.6.1 Parameter Settings

The default settings of Random Forest and Naïve Bayes were not changed. In Chapter Three we described the SVM algorithm and the parameters that require changing in order for it to be adaptable to the domain and the underlying dataset. It could be recalled that the parameter *C* represents data complexity which regulates the bias and variance properties as observed in the data. Several values were experimented with, taking into account the potential variance problems associated with the VA dataset. The values between 1.0 and 0.01 were experimented with in turn, with 0.05 being found to be the best performing parameter setting for this dataset. This value was observed to achieve optimum performance, which confirmed our observation of the high variance associated with the VA domain.

---

[1] Attribute-Relation-File-Format

7.6.2 Evaluation Method

The experiments employed the 10 fold cross validation evaluation method to allow a random stratification by the categories into training and test sets of 10 folds (Kohavi, 1995). A weighted average was then computed over the 10 folds as discussed in Chapter Four. We employed *macro–averaging* to determine the overall performance due to the highly uneven distribution of the multi-class dataset being used, which allowed equal weights to be computed for each cause of death category (Forman, 2004). We also reported the percentage average overall accuracy discussed in Chapter four, so that it could be compared against a majority baseline which was pre-determined.

7.6.3 Statistical Significance Testing

We employed the *Fisher's randomisation test* (Box et al., 1979) to test for significance difference between two competing classifiers. Our choice was based on the underlying non-normal distribution assumption of this test, which made it the best for our data due to the imbalanced characteristic found in the VA corpus as already demonstrated in Chapter Six. To achieve this, we adapted the randomisation test procedure described by Smucker et al.(2007) and implemented it using the StatKey v. 0.3.12[2] software. Our procedure however varied in two respects: we considered difference in percentage overall average accuracies to be able to determine any effect in performance, even in situations where precision values remained the same. Changes in percentage overall average accuracy and macro-average f-measure values were also observed due to the imbalance issues associated with our corpus. It must be noted that Smucker et al.(2007) pointed out that the use of an alternative performance metric such as accuracy or f-measure has no difference in its method and therefore yields similar outcomes and conclusions. We also generated 20,000 random samples instead of the 100,000 used in Smucker et al.(2007)'s experiment. This was done after observing no change in values upon experimenting with trials between 1,000 and 100,000 with a step of 5,000 and no change in p-values being observed. This was found to be consistent with observations by Jensen and Cohen (2000), who found a random resample of 1000 should be enough to observe any difference between algorithms. A right-tail test was used to determine the p-value (p) with a 95% confidence interval (CI= 95%) in order to reject the null

---

[2] http://lock5stat.com/statkey/index.html

hypothesis. A screen shot of a sample output from the StatKey significance test can be found in appendix C.

7.7 Results

7.7.1 Baseline

As indicated, it is important to establish a baseline for the experiments and for the various groupings.

Table 7.1a Time-of-Death results: baseline for Groupings1

Percentage accuracy	macro-average-f-measure	majority category
31.3	0.14	Neonatal

Table 7.1b Type-of-Death results: baseline for Groupings1

Percentage accuracy	macro-average-f-measure	majority category
22.1	0.08	Stillbirth_unexplain

Table 7.1c Time-of-Death results: baseline for Groupings2

Percentage accuracy	macro-average-f-measure	majority category
35.7	0.18	Neonatal

Table 7.1d Type-of-Death results: baseline for Groupings2

Percentage accuracy	macro-average-f-measure	majority category
25.2	0.10	Stillbirth_unexplain

Table 7.1e Time-of-Death results: baseline for Groupings3

Percentage accuracy	macro-average-f-measure	majority category
42.3	0.25	Stillbirth

Table 7.1f Type-of-Death results: baseline for Groupings3

Percentage accuracy	macro-average-f-measure	majority category
24.5	0.09	Antepartum_stillbirth

Tables 7.1a and 7.1b show the baseline results obtained for *Groupings1 Time-of-Death* and *Type-of-Death*. As the tables show, an overall accuracy of 31.3% (Neonatal) and 22.1% (Stillbirth_unexplain) with macro-average f-measure score of 0.14 and 0.08 were obtained for *Time-of-Death* and *Type-of-Death* respectively. Furthermore, tables 7.1c and 7.1d show the results obtained for *Groupings2*: For *Time-of-Death*, an overall accuracy of 35.7% (Neonatal) with a macro-average f-measure score of 0.18; Type-of-Death overall accuracy of 25.2% (Stillbirth_unexplain) with a macro-average f-measure score of 0.10. Finally, tables 7.1e and 7.1f show the baseline results obtained for *Groupings3*: *Time-of-Death*, an overall accuracy of 42.3% (Stillbirth) with a macro-average f-measure score of 0.25; *Type-of-Death* overall accuracy of 24.5% (Antepartum_stillbirth) with a macro-average f-measure score of 0.09.

7.7.2 Feature Value Representation Schemes and Classification Algorithms

Tables 7.2a and 7.2b show the variations that existed in the performance of feature value representations. The *Binary* scheme achieved the worst performance across all the three learning algorithms. This was followed by the *Term Frequency* scheme. *Normalised Term Frequency,* was demonstrated to be superior to *TFiD* across all algorithms for *Time-of-Death*: NB (accuracy: 4.1% (43.4 - 39.3); macro-average-f-measure: 0.5 (0.42 - 0.37); p<0.001); RF (accuracy: 2.4% (33.7 - 31.3); macro-average-f-measure: 0.4(0.41 - 0.37), p=0.001) and SVM (accuracy: 1.3% (45.0 - 43.7), macro-average-f-measure: 0.1(0.42 - 0.41); p=0.06). Similarly, as shown in table 7.2b *Type-of-Death* categories, *Binary* is the least performing representation scheme, followed by the *Term Frequency*. The *Normalised Term Frequency* outperformed *TFiD* across all algorithms for *Type-of-Death*: NB (accuracy: 1.3% (26.2 - 25.1); macro-average-f-measure: 0.1 (0.24 - 0.23), p=0.09); RF(accuracy: 0.1% (22.2 - 22.1), macro-average-f-measure: 0 (0.8 - 0.8), p=0.024) and SVM (accuracy: 1.4% (30.1 - 26.2), macro-average-f-measure: 0.01(0.25 - 0.23), p<0.001).

Comparing classification algorithms based on the best performing feature value representation (*Normalised Term Frequency*), *Random Forest* had the worst performance. Thus, we compare SVM and Naïve Bayes. *Time-of-Death* (accuracy: 2.1% (45.0 - 43.4), macro-average-f-measure: 0.1(0.42 - 0.41); p<0.05). *Type-of-Death* (accuracy: 1.1% (26.2 - 25.1); macro-average-f-measure: 0.1 (0.24 - 0.23), p<0.001). SVM significantly outperformed Naïve Bayes as the p-values suggest.

Table 7.2a Time-of-Death results: performance comparison between feature value representations and classification algorithms.

	Term Frequency		Binary		Normalised Term Frequency		TFiDF	
algorithm	percentage accuracy	macro-average-f-measure	percentage accuracy	macro-average-f-measure	percentage accuracy	macro-average-f-measure	percentage accuracy	macro-average-f-measure
Random Forest(RF)	31.3	0.15	31.3	0.15	33.7	0.20	31.3	0.15
Naïve Bayes(NB)	36.7	0.36	33.5	0.25	43.4	0.41	39.3	0.37
SVM	42.1	0.39	31.3	0.15	45.0	0.42	43.7	0.41

Table 7.2b Type-of-Death results: performance comparison between feature value representations and classification algorithms.

	Term Frequency		Binary		Normalised Term Frequency		TFiDF	
algorithm	percentage accuracy	macro-average-f-measure	percentage accuracy	macro-average-f-measure	percentage accuracy	macro-average-f-measure	percentage accuracy	macro-average-f-measure
Random Forest(RF)	22.1	0.08	22.1	0.08	22.2	0.08	22.1	0.08
Naïve Bayes(NB)	24.8	0.22	24.4	0.19	26.2	0.24	25.1	0.23
SVM	28.6	0.23	22.1	0.08	30.1	0.25	28.7	0.24

7.7.3 Feature Reduction as a Cure to High Dimensionality: Performance Comparisons.

Having identified SVM as the suitable algorithm for this task, our subsequent experiments considered only SVM. Tables 7.3a and 7.3b show the results obtained from the various feature reduction schemes employed. The *locally-semi-automatic* approach describes the results obtained from various thresholds explored, which started with the top-ten words of each of the *Time-of-Death* and *Type-of-Death* cause of death categories as per the log-likelihood rankings described in the procedure above. The top-250 words achieved the highest performance.

Table 7.3a Time-of-Death results: feature reduction strategy

feature reduction strategy	percentage accuracy	macro-average-f-measure
SVM(unigram)	45.0	0.42
locally-semi-automatic	45.8	0.43
Information Gain	39.2	0.32
String-matching	31.2	0.15

Table 7.3b Type-of-Death results: feature reduction strategy

feature reduction strategy	percentage accuracy	macro-average-f-measure
SVM(unigram)	30.1	0.25
locally-semi-automatic	30.3	0.25
Information Gain	23.4	0.11
String-matching	22.1	0.08

Tables 7.3a and 7.3b show the performance of the feature reduction strategies employed for *Time-of-Death* and *Type-of-Death* categories respectively. As seen from the table, the *String-matching* algorithm is the least performing strategy, achieving performance even below the SVM (unigram): there was a significant drop in performance for both the *Time-of-Death* (accuracy: -13% (31.2 - 45.0); macro-average-f-measure: -0.27(0.15 - 0.42) ; $p<0.001$) and the *Type-of-Death* (accuracy: -8% (22.1 - 30.1); macro-average-f-measure: -

0.17(0.08- 0.25); p<0.001). *Information Gain* on the other hand significantly outperformed *String-matching at* both *Time-of-Death* (accuracy: 8% (39.2 - 31.2); macro-average-f-measure: 0.27(0.32 - 0.15); p<0.001) and *Type-of-Death* (accuracy: 1.3% (23.4 - 22.1); macro-average-f-measure: 0.03(0.11- 0.08); p<0.03). It was however surprising to see a significant drop in performance when comparing *Information Gain* to SVM (unigrams) for both *Time-of-Death* (accuracy: -5.3% (39.2 - 45.0); macro-average-f-measure: -0.10 (0.32 - 0.42); p<0.001) and *Type-of-Death* (accuracy: -6.7% (23.4 - 30.1); macro-average-f-measure: -0.14 (0.11- 0.25); p<0.001). The results further suggested a marginal increase (not statistically significant) in performance accuracy when comparing the *locally-semi-automatic* to SVM (unigrams) for both *Time-of-Death* (accuracy: 0.8%(45.8 - 45.0); macro-average-f-measure: -0.10(0.43- 0.42); p=0.18) and *Type-of-Death* (accuracy: 0.2%(30.3 - 30.1); macro-average-f-measure: 0 (0.25- 0.25); p=0.4). A significant increase in performance was observed when comparing *locally-semi-automatic* to *Information Gain* for both *Time-of-Death* (accuracy: 6.6% (45.8 - 39.2); macro-average-f-measure: 0.11(0.43 - 0.32); p<0.001) and *Type-of-Death* (accuracy: 6.9% (30.3- 23.4); macro-average-f-measure: 0.14(0.25- 0.11); p<0.001).

7.8 Discussion

Feature value representation and classification

Although *Random Forest* has successfully been applied to classify the closed response part of VA data (Flaxman et al., 2011), the results obtained from this experiment suggest that it is not an appropriate choice for classification of VA open narrative text. This may be due to the differences that exist between the feature vectors generated from the closed and the open narrative data may account for this. The closed response part data feature vector is derived from a controlled vocabulary with limited number of features, possibly a list of questions with *yes* and *no* answer options Figure 1.1 shows. In contrast, the uncontrolled vocabulary characteristic of VA open narrative text results in a large number of features. The relatively poor performance of Random Forest in this experiment could be explained by the deficiency which have been found to be generally associated with Decision Trees based algorithms which include Random Forest as being susceptible to over-fitting with training data of 500 or more features (Joachims, 1998). This may have harmed the *Random Forest* algorithm since the text tends to generate a high number of features. The closed response part of VAs is unlikely to exceed the 500 features limit, whereas 12,304 features(unigram) was used in this experiment and may therefore be unsuitable for the *Random Forest* learning algorithm.

Also, the independent assumption applied in *Naïve Bayes* may explain the relatively better performance compared with *Random Forest*, and even its performing better than *SVM* for the *Binary* representation scheme. This is because *Naïve Bayes* was noted to be highly sensitive to majority category, as it has tended to be susceptible to imbalanced data, which resulted in it achieving a relatively better overall accuracy for skewed data (Rennie et al., 2003). Furthermore, *Binary* representation may not contain enough useful information for SVM to learn from.

As noted from the results, with the exception of the *Binary* representation, the consistent superior performance of the SVM algorithm to both *Naïve Bayes* and *Random Forest* algorithms was not surprising. SVM has been consistently shown to have a relatively better performance in Text Classification experiments (Witten and Frank, 2005), and the results from this experiment were not an exception. The outstanding performance of SVM could be attributed to a number of factors: the majority of Text Classification problems are mostly linearly separable, and SVM employs threshold functions to develop margins that linearly separate the categories; SVMs also tend to use an over-fitting protection mechanism that is independent of the dimensionality of the feature space, thus, the number of features tends not to be an issue; and SVMs are well designed to deal with sparseness found in feature vectors. The Danso et al.(2013a)'s description of the VA text correlates with the taxonomy of issues outlined that the SVM algorithm was designed to address. It therefore seems natural that SVM tends to perform better than *Naïve Bayes* and *Random Forest* for this task.

Feature reduction strategy

The *String-matching* strategy was also observed to have achieved performance equivalent to the *Baseline* algorithm (see table 7.1a and 7.1b). The reasons for this poor performance could be summarised as:

- Less usage of medical terms*:* although the terms used in describing symptoms were known to be basic medical terms as contained in the lexicon created from the online "kids" medical thesaurus, the experiments suggest that there were not enough of these terms. As indicated in Chapter Six, local terms and non-standard medical terms were predominantly used in expressing medical concepts. The poor results obtained from these experiments are a confirmation of this.

- Variations in spellings of medical terms*:* it is however worth noting that some variations in spelling of medical terms were observed as being results of misspelling and so were not taken into account for the *String-matching* strategy. Normalisation of these variations should improve on this current result but may not be significant enough to make a difference due to the reasons given above.

Furthermore, as the tables demonstrate the *Information Gain* algorithm outperformed the *String-matching* strategy. It was however surprising to observe that *Information Gain* performed worst against the SVM (unigram). The reason that might have accounted for this result was that the *Information Gain* employed a global feature reduction technique in determining the most informative terms for all categories (Forman, 2003). This approach may not have been effective if there were overlapping terms between the categories. For example, the term *dead* may be found in all categories, which is not informative enough to differentiate between *intra-partum* and *ante-partum stillbirth* cause of death classes.

Moreover, when comparing the *locally-semi-automatic* with *Information Gain,* there was a significant difference in performance in favour of *locally-semi-automatic*. One possible reason for the *locally-semi-automatic* approach outperforming the *Information Gain* is due to the fact that unlike the *Information Gain* approach, the terms obtained by the *locally-semi-automatic* approach were locally weighted against all other terms found in the corpus, which allowed the most informative terms for each category to be identified. Other terms introduced noise which potentially degraded the performance of the algorithms. Despite the strength of this approach, the disadvantage of it is that it is error prone. This may be due to the manual selection iteration of the various thresholds required by the process in order to identify the threshold required to achieve optimal performance. Therefore, a full automation of the process would be beneficial.

SVM's robustness to over-fitting has resulted in the argument within the research community that it is irrelevant to carry out feature reduction before algorithm training (Forman, 2003). The obtained from this experiment supports this claim given the fact that no significant difference between the SVM (unigram) and the *locally-semi-automatic* which is the best performing feature reduction strategy among the rest considered in these experiments.. We would however argue that there is some additional benefit in reducing features as a prior step to performing learning, though this does not lead to a reduction in performance. The computational demand required in processing can be significantly reduced by cutting down the number of features (unigram). Only 12% (1000 out of over 12,300 terms of the unigram) of features achieved the performance obtained by the *locally-semi-automatic* approach. This approach to feature reduction may have effectively selected features that have stronger correlation with the cause of death categories.

7.9 Summary

This Chapter presented results from experiments carried out to explore various techniques suitable for the classification of VA open narrative text: feature value representation; Machine Learning algorithms and feature reduction strategies. Our experimental results suggested that *Normalised Term Frequency* has slightly better performance over *TFiDF,* and significantly better than *Binary* and *Term Frequency*. In terms of Machine Learning algorithms, the SVM algorithm was found to be the best performing algorithm and the most suitable for the VA domain. Furthermore, this Chapter proposed a feature reduction method known as *locally-semi-automatic*, which was based on a log-likelihood statistical measure, and a comparison was made to determine the effectiveness of this method compared with other standard methods. The experimental results also show that employing a *locally-semi-automatic* method to identify informative features resulted in a substantial improvement in accuracy when compared against other standard feature reduction approaches.

Chapter 8

Linguistics and Statistically derived features for Time and Cause of Death prediction from Verbal Autopsy narratives

"the true method of knowledge is experiment", William Blake

8.1 Introduction

In Chapter Seven we explored various methods and techniques that formed the foundations of the classification of VA text. This included how features should be represented, with the results suggesting *Normalised Term Frequency* as the best feature value representation scheme for our task. We also established SVM as the most suitable Machine Learning algorithm for this task. We further explored various feature reduction strategies and demonstrated that our method, known as the *locally-semi-automatic* approach, can be used to effectively determine the most informative words out of all available words, which we called the *Discriminative Word Units* (DWU) for developing models for Text Classification. We demonstrated that DWU achieved a marginally better performance than unigrams and a significantly better one than a baseline. We explained that this approach was particularly useful when dealing with a very noisy text such as VA.

In this Chapter, we expand on the results obtained from Chapter Seven by describing further experiments that we carried out. The aim of these was to further explore the feature space in order to improve on the results obtained in Chapter Seven. The various features that were explored in the experiments and the motivation for their use is further discussed here. We refer to the model developed from the experiments described in this Chapter as VAModel1. We conclude by discussing the behaviours, contributions and effects of these features on the performance of our models.

8.2 Linguistic Features

Linguistic features are considered features with some grammatical information that can be derived from the context in which the words occur. We employed Part-of-Speech (PoS) tagging to obtain this information. This approach was considered to be a crude form of determining the correct sense of a given word in a text (Wilks and Stevenson, 1998). In Chapter Six we demonstrated the performance of three PoS taggers including Brill Transformation-based Learning algorithm (Brill, 1995) available as part of the NLTK (Loper and Bird, 2002). Furthermore we established the fact that the standard PoS tagging approaches available have shown to achieve comparable results on a VA corpus. Thus, the Brill tagger was trained using the Brown corpus (Francis and Kucera, 1979) to tag words in the VA open narrative document. It is important to note that the tagging was done on all words found in each VA document and not just the DWU.

8.2.1 Part-of-Speech Tag Patterns

PoS tag patterns are linguistic information obtained from PoS taggers. PoS tags have been shown to be a useful feature in numerous Text Classification problems. Gamon (2004), for example, demonstrated the use of PoS trigrams in sentiment classification. We explored various PoS tags in our experiments which included unigram, bigrams and trigrams, but no significant difference was observed between the various approaches. This was due to the nature of the VA text as we have demonstrated in Chapter Five and also discussed in Danso et al. (2013a) and Danso et al. (2013b). The results presented in this thesis are therefore based on PoS unigram.

8.2.2 Noun and Verb Phrase Tag Patterns

Having obtained PoS information for every word in the text, a chunking technique implemented by the regular expression below were used to extract noun and verb phrases. The motivation to explore these features was inspired by the fact that domain concepts are mostly expressed using multiword structures (Moschitti and Basili, 2004). Furthermore, chunking allows PoS tags to be extracted from ill-formed sentences through shallow parsing (Clark, 2003). Once chunking had been done, the PoS tags enabled us to extract various tag sequences that form noun and verb phrases.

```
r'''NounPhrase:{<DT>?<JJ>*<NN>+}
               a  normal labour
```

Figure 8.1 regular expression to extract noun phrase

For example, as demonstrated in Figure 8.1, the pattern captured a multi-word concept *a normal labour* used to describe a type of labour a mother experienced during pregnancy, which was domain specific information. Similarly, because verb phrases tend to capture descriptive information about some action or inaction in a sentence, they tend to feature frequently in the VA open narrative to describe a condition due to illness. This is illustrated by the example in Figure 8.2. The Figure shows how the concept *I lost appetite* is extracted in the text.

```
r'''VerbPhrase:{<PPSS>?<VBD>+<NN>+}
                 ↓        ↓     ↓
                 I      lost  appetite
```

Figure 8.2 regular expression to extract verb phrase

A generalised approach to capturing these types of mentions in the text is through extraction of these phrases, which are derived from the PoS tags. Another reason is that this approach tends to be a convenient method of extracting semantic information from the text without being overdependent on the actual words that appear in the running text due to variations and errors in the spellings, which results in further data sparseness problems. The Noun and Verb phrase tag sets were represented as single items as part of the feature set for modelling. As Figure 8.2 demonstrates, our verb chunking algorithm considered only past tenses, the reason being that a review of VA sample documents suggested tenses were mainly expressed in the past, which supported the intuition that VAs were concerned with past events. It is therefore not surprising that Smith et al.(2014) also found a similar pattern in the Swedish clinical narrative text.

8.3 Lexical Features

Lexical features have been defined in this thesis as information that can be derived from the individual words found in the VA open narrative text. We will now describe some of the lexical-based features and how these were extracted.

8.3.1 Statistically Derived Features

Statistically derived features are considered to describe some form of phenomena that tend to occur in the use of a language but are not predictable. As observed by Harris (1951), "*each word has a particular and roughly stable likelihood of occurring as argument, or operator, with a given word, though there are many cases of uncertainty disagreement among speakers, and change through time*". Collocation extraction is one technique that can be used to capture the phenomena described above. Collocation extraction could be considered a lexical-based method since it is employed to capture word-pairs and phrases that frequently occur in the use of a language with no regard to their semantic or syntactic rules of use. The extracted word-pairs and phrases are also known to be dialect or language specific (McKeown and Radev, 2000).

Collocation extraction-based phrases

Collocation extraction has been employed in many applications by lexicographers to carry out word sense disambiguation and semantic analysis of text (Pearce , 2001).This therefore suggests an imperative investigation should be done into the potential use of collocation as a feature to identify patterns of co-occurrence of words that could be indicative of a phrase or expression of cause of death, considering the peculiar nature of colloquial text contained in the VA corpus.

We have employed statistical methods based on log-likelihood estimation to determine the likelihood of co-occurrence words and phrases (Dunning, 1993). The log-likelihood estimation was based on the entire corpus and estimated the likelihood of two words co-occurring as defined by the bigram log-likelihood statistics association measure (Pearce, 2002) to retrieve words from the corpus that tend to collocate with DWU subset of words already obtained, as described in Chapter Seven. The limitation associated with the bigram approach in general is that it usually takes into account the only two word-collocates (bigrams) (Seretan et al., 2003), which introduces data sparseness into the feature space. In order to reduce the impact of this and also maximise the additional information that can be derived from the collocated words, we explored levels of association observed from the corpus as ranked by the bigram log-likelihood statistics association measure algorithm. We subsequently retrieved words with the strongest association with words in our DWU list. This means that the words in the corpus that collocate most with the DWU words were retrieved and experimented in turns and their impacts on performance were obtained. We illustrate our idea with the following example.

during | 'labour'= 4150,'pregnancy'= 2901 , 'my' = 1785

Figure 8.3 an example of collocation extraction-based phrase

Figure 8.3 shows an example with the word *during* which is part of the DWU list. This was used to retrieve the top-ranked words with the strongest association and their corresponding likelihood values, as ranked by the bigram log-likelihood statistics association measure (Dunning, 1993). As the Figure demonstrates, *during* collocated with *labour* more frequently and this was followed by *pregnancy,* before *my,* with their respective log-likelihood values.

8.3.2 Simplified Relative Word Position

Simplified relative word position is defined as the position of a given word as it appeared in the document. Our motivation for experimenting with this as a possible feature was based on one of the criticisms put forth of the bag-of-words approach to Text Classification, which was that the bag-of-words approach ignored the order and structure in sentence (Scott and Matwin, 1999).The idea of exploration of positions of words in text as possible features for Text Classification has been explored by various researchers (Matsumoto et al., 2005, Pang and Lee, 2004). Matsumoto et al.(2005), for example, demonstrated the usefulness of this feature by extracting word sub-sequences and dependency sub-trees from sentences to classify movie reviews. We however adopted a simplified approach by exploring the relative positions of the words as they appeared in the text to capture the sequential order of events within the context of the VAs.

The approach employed in this thesis considered the entire content of a VA document as a single string of words with an imaginary grid, where each cell represented a word which was a member of the string. Each cell was serially allocated a unique number and that represented the position of the word with respect to the entire string. The position number of the word captured was divided by the length of the string (number of cells) to obtain its relative word position in the VA document.

The hypothesis suggested was that there may be a logical order of events in the history of an individual which led to their death, and that might be a major factor in case profiling or an investigation process. This feature could help in capturing that order. To effectively illustrate this phenomenon, we considered an example as shown in Figure 8.4a and 8.4b where sentences were taken from a VA document.

D_1 = "*In the second month of the pregnancy,…labour started which I was at home in the morning…*"

Figure 8.4a sample sentence extracted from a VA document

D_2 = "*In the second month labour started….*"

Figure 8.4b hypothetical case of ignored word order in a VA document

If the order of these words were to be ignored one possible reading could be the example shown in Figure 8.4b, document D_2 which presented a different scenario and might have meant a different outcome which could have compromised the medical perspective. The proposed *simplified relative word position* feature aimed to avoid such a situation and preserved the order in which the words appeared in the VA document as illustrated in Figure 8.5.

Term	D_1	D_2
In	p_1t_1	p_1t_1
the	p_2t_2	p_2t_2
second	p_3t_3	p_3t_3
month	p_4t_4	p_4t_4
of	p_5t_5	p_5t_5
pregnancy	p_6t_6	
labour		p_6t_6
…		

Figure 8.5 feature vector representation of documents D_1 and D_2 with relative word position feature encoded

The resulting feature vector as shown in Figure 8.5 would therefore be encoded as part of the feature value for that vector, which would result in different term weights ($p_1t_1..p_6t_6$) for document D_1 and D_2, where *p* = position and *t*= normalised term frequency.

8.4 Cause of Death Regroupings

It will be recalled in Chapter Six that three causes of death groupings were proposed. Again, in Chapter Four, the effect of category noise and the importance to training set with non-overlapping categories were discussed. This batch of experiments explored the effect of the regroupings on performance of our model.

8.5 Results

The meaning of a word is best known by the context in which it exists (Firth, 1957) and this was evident in the results obtained from these experiments. Using various linguistic and statistically derived features, which have extra information about the individual words, we were able to achieve a significant increase in the performance accuracy over the single words - the DWU in predicting cause of death. The effect of each of these features will be examined further now.

8.5.1 The Effect of Part-of-Speech Tags

We examined the effect of introducing PoS-tags as features.

Table 8.1a Time-of-Death results: effect of PoS-Tags

features	percentage accuracy	macro-average-f-measure
DWU	45.8	0.43
DWU+ PoS-Tags	46.8	0.44

Table 8.1b Type-of-Death results: effect of PoS-Tags

features	percentage accuracy	macro-average-f-measure
DWU	30.3	0.25
DWU + PoS-Tags	31.3	0.27

Tables 8.1a and 8.1b show the performance obtained when the PoS-Tags feature was added to the DWU feature set: *Time-of-Death* (accuracy: 1% (46.8 - 45.8); macro-average-f-measure: 0.01 (0.44 - 0.43); p=0.3) and *Type-of-Death* (accuracy: 1% (31.3 - 30.3); macro-average-f-measure: 0.02 (0.27 - 0.25); p=0.1). As the figures suggest, there was a marginal increase in performance but not one statistically significant as the p-value suggests.

8.5.2 The effect of Noun and Verb phrases

We examined the effect of Noun and Verb phrases on the overall performance of our model when added as a feature to the DWU + PoS-Tags feature set

Table 8.2a: Time-of-Death results: effect of NounAndVerbPhrase

features	percentage accuracy	macro-average -f-measure
DWU + PoS Tags	46.8	0.44
DWU + PoS-Tags + NounAndVerbPhrase	47.0	0.45

Table 8.2b: Type-of-Death results: effect of NounAndVerbPhrase

features	percentage accuracy	macro-average -f-measure
DWU + PoS Tags	31.3	0.27
DWU + PoS-Tags + NounAndVerbPhrase	31.3	0.27

Tables 8.2a and 8.2b contain the results obtained from the introduction of the NounAndVerbPhrase feature. As the results clearly demonstrate, there was no significant change in the results when compared against the results obtained prior to its addition to the DWU+PoS-Tags features. *Time-of-Death* (accuracy: 0.2% (47.0 - 46.8); macro-average-f-measure: 0.01 (0.45 - 0.44); p=0.47). No change in performance was observed for the *Type-of-Death* category. Also, even though a slight increase in performance was observed for the Time-of-Death category, this was highly insignificant as the p-value suggests.

8.5.3 The Effect of Simplified Relative Word Positions

We determined the effect of the *simplified relative word position* feature on the overall performance of our model.

Table 8.3a Time-of-Death results: effect of simplified relative word position

features	percentage accuracy	macro-average-f-measure
DWU + PoS-Tags + NounAndVerbPhrase	47.0	0.45
DWU + PoS-Tags + NounAndVerbPhrase + Simplified-relative-word-position	56.2	0.55

Table 8.3b Type-of-Death results: effect of simplified relative word position

features	percentage accuracy	macro-average-f-measure
DWU + PoS Tags + NounAndVerbPhrase	31.3	0.27
DWU + PoS-Tags + NounAndVerbPhrase + Simplified-relative-word-position	36.4	0.29

Tables 8.3a and 8.3b show the results obtained after the introduction of the *simplified relative word position* feature to the *DWU+PoS Tags+NounAndVerbPhrase* feature set. *Time-of-Death* (accuracy: 9.2% (56.2 - 47.0); macro-average-f-measure: 0.10 (0.55 - 0.45) ; p<0.001) and *Type-of-Death* (accuracy: 5.1% (36.4- 31.3); macro-average-f-measure: 0.02 (0.29 - 0.27); p<0.001). As the values suggest, there was a significant increase in performance for both *Time-of-Death* and *Type-of-Death* categories.

8.5.4 The Effect of Collocation Extraction

It was hypothesised that collocation based phrase extraction should enable the detection of phrases that are particularly used by respondents in describing symptoms and events of death. We investigated the effect of the various levels as ranked by the collocation extraction algorithm on performance.

Table 8.4a Time-of-Death results: effect of top-three collocates

features	percentage accuracy	macro-average-f-measure
DWU + PoS-Tags + NounAndVerbPhrase + Simplified-Relative-Word-Position	56.2	0.55
DWU + PoS+Tags + NounAndVerbPhrase + Simplified-Relative-Word-Position + Top-One-Collocate	56.8	0.55
DWU + PoS+Tags + NounAndVerbPhrase + Simplified-Relative-Word-Position + Top-One-Collocate + Top-Two-Collocate	57.5	0.56
DWU + PoS Tags + NounAndVerbPhrase + Relative-Word-Position + Top-Three-Collocate	59.4	0.57

Table 8.4b Type-of-Death results: effect of top-three collocates

feature	percentage accuracy	macro-average-f-measure
DWU + PoS-Tags + NounAndVerbPhrase + Relative-Word-Position	36.4	0.29
DWU+PoS-Tags + NounAndVerbPhrase + Relative-Word-Position + Top-One-Collocate	36.5	0.29
DWU + PoS-Tags + NounAndVerbPhrase + Relative-Word-Position + Top-Two-Collocate	37.2	0.31
DWU + PoS Tags + NounAndVerbPhrase + Relative-Word-Position + Top-Three-Collocate	38.0	0.36

Tables 8.4a and 8.4b show the breakdown of the ranked collocates for *Time-of-Death* and *Type-of-Death* categories respectively prior to and after the introduction of the collocation extraction based feature: *Time-of-Death* (accuracy: 3.2% (59.4 - 56.2); macro-average-f-measure: 0.02 (0.57 - 0.55); p<0.001). *Type-of-Death* (accuracy 3.2% (38.0 - 35.4); macro-average-f-measure: 0.02 (0.57 - 0.55); p=0.009). The figures show a significant increase in performance for both *Time-of-Death and* Type*-of-Death.* Also, in terms of breakdown, the tables show the contribution of the various best-ranked collocates and their effect on performance. A gradual progression of increase

in performance was observed from the introduction of the first to the third topmost collocates included in the feature set, which resulted in a significant overall increase in performance as demonstrated.

8.5.5 The Effect of Cause of Death Category Regroupings

This section examines the effect of re-grouping cause of death classes on performance of the classification models. This is based on all the features explored above.

Table 8.5a Time-of-Death results: effect of category regroupings

grouping	percentage accuracy	macro-average f-measure
Groupings1	59.4	0.54
Groupings2	65.6	0.65
Groupings3	85.4	0.85

Table 8.5b Type-of-Death results: effect of category regroupings

grouping	percentage accuracy	macro-average f-measure
Groupings1	38.0	0.36
Groupings2	40.1	0.37
Groupings3	46.3	0.39

Tables 8.5a and 8.5b are results obtained from exploration of the impact of re-grouping cause of death classes on performance of the classification models. As the tables demonstrate, significant improvements in performances were observed for both *Time-of-Death* and *Type-of-Death*. An initial increase in performance was observed between *Groupings2* and *Groupings1*: *Time-of-Death* (accuracy: 6.2% (65.6 - 59.4); macro-average-f-measure: 0.11(0.65 - 0.54); $p<0.0001$). *Type-of-Death* (accuracy: 2.1% (40.1 - 38.0); macro-average-f-measure: 0.01(0.37 - 0.36); $p=0.008$). A further significant increase in performance was observed when the performances of *Groupings3* were compared against *Groupings1: Time-of-Death (*accuracy: 26% (85.4 - 59.4); macro-average-f-measure: 0.31(0.85 - 0.54); $p<0.001$) and *Type-of-Death* (accuracy: 8.3% (46.3 - 38.0); macro-average-f-measure: 0.3 (0.39 - 0.36); $p<0.001$). Also, a significant difference in performance was observed between

Groupings3 and *Groupings1*: *Time-of-Death (*accuracy: 19.8% (85.4 - 65.6); macro-average-f-measure: 0.30 (0.85 -0.65); p<0.001) and *Type-of-Death* (accuracy: 6.2% (46.3 - 40.1); macro-average-f-measure: 0.02 (0.39 - 0.37); p<0.001). This suggested that *Groupings3* achieved the best performance.

8.5.6 Overall Performance Achieved for Open Narrative Text

In order to assess the overall performance accuracy obtained between the *Time-of-Death* and *Type-of-Death* levels of classification, we compared the performance difference between the baseline results and our best performing classification models (*Groupings3*).

Table 8.6 demonstrates a significant increase in performance when compared with the performance obtained at baseline for the same groupings. *Time-of-Death* (accuracy: 43.1% (85.4 - 42.3); macro-average-f-measure: 0.29 (0.25 - 0.54); p=0.01) and *Type-of-Death* (accuracy: 21.8% (46.3 - 24.5); macro-average-f-measure: 0.31(0.39 - 0.08) p =0.007).

Table 8.6 Overall performance comparisons: baseline vs. achieved for open narrative text

Level	percentage accuracy			macro-average-f-measure		
	baseline	achieved	difference	baseline	achieved	difference
Time-of-Death	42.3	85.4	43.1	0.25	0.54	0.29
Type-of-Death	24.5	46.3	21.8	0.09	0.39	0.30

8.6 Discussion

The above results presented an interesting insight into the behaviour of the features and the amount of information they tended to contribute to the overall process of prediction. It can be inferred from the results that the most substantial increase in performance of the classifier was due to the *simplified relative word position* feature which tends to support the underlying intuition of our approach to generating these features as discussed earlier. The feature tended to effectively capture the patterns in the usage of words. It also suggested a certain pattern in the order in which these words are used in narrating the history of events. This feature appears to have an accurate representation and also mimics how physicians tend to navigate through the text when building up a mental picture of a given history in order to arrive at a judgement of the possible cause of death.

It is important to note, however that this was a variation of the relative word position based features proposed in the literature for modelling Text Classification (Kudo and Matsumoto, 2004; Matsumoto et al., 2005) as previously mentioned. Those approaches required deep parsing and syntactic analysis of the sentence structure, which may also have thus required well-formed sentences. However, our simplified approach did not need such detailed analyses of the sentences. We demonstrated that it was possible to represent the entire content of the document as a single string, and this presents useful information for Machine Learning algorithms to learn patterns from in the future. This was a particularly useful feature for ill-formed text in documents where deep linguistic analysis could not be achieved or remained problematic as demonstrated from the results of our experiments.

The results shown above also demonstrated the usefulness of the topmost 3 collocation extraction based feature. As hypothesised from the rationale and motivation for generating these features, this method enabled us to successfully identify and extract phrases that were indicative of the various causes of death, which resulted in performance improvements. Another significant characteristic observed about this feature was the relatively higher precision obtained compared to the corresponding recall and, consequently, narrowing the gap between recall and precision. Based on these findings it could be argued that the ranking and retrieving of the 3 topmost collocates tended to capture a *psycholinguistic* phenomenon. Our results demonstrated the presence of this phenomenon in narrating causes of death as captured in the VA open narrative text.

The relatively low improvements observed from the PoS tag patterns that were derived from linguistic analysis of the text demonstrated that the information obtained from PoS tags tended to contribute information required by a classifier. However, considering our hypothesis about the potential usefulness of noun and verb phrases in capturing multiword concepts, it was disappointing to observe no improvement in performance from the introduction of noun and verb phrase tag pattern feature set. However, the results obtained from the noun and verb phrase-based features tended to be consistent with other findings (Scott and Matwin, 1999) and so this area requires further investigations.

Considering the noisy and rather unusual type of text being dealt with, there was the possibility that the features employed so far may have not been effective enough in discriminating between all causes of death. There is therefore the need for further exploration within the feature space of the narratives in order to improve on the performance obtained. This would include adaptation of the standard PoS taggers for this particular type of text. This was demonstrated by the PoS tagging experimental results obtained from the PoS tagger evaluation experiment we conducted (Danso et al., 2013a). It may be argued that the choice of the Brown corpus for training the PoS tagger was inappropriate considering the difference in text, which may have resulted in the poor performance of the linguistic features extracted from the output of the PoS tagger. It must however be pointed out that the choice was purely based on convenience as there was no linguistically annotated corpus readily available similar to the VA corpus. The possibilities of training the PoS tagger with a corpus that has linguistic annotations from either the VA or biomedical domain should be explored in order to observe the impact of these features on the overall performance of our model. For example, it would be interesting to see how the PoS tagger developed by Tsuruoka et al.(2005) for biomedical text performs on VA text.

Several factors may account for the differences in performance accuracies observed between the two levels of classification: *Groupings3*, which was the best performing model, had three and 10 categories for *Time-of-Death* and *Type-of-Death* respectively. Consequently, the amount of training data for the *Time-of-Death* category was bigger compared to the *Type-of-Death* category. As Banko and Brill (2001) demonstrated in their experiment on the effect of size of training data on performance of a Machine Learning algorithm, their experiment confirmed this phenomenon. Again, as discussed in Chapter Four, with regards to issues that tend to impact on Machine Learning algorithms, the prediction at the fined-grained (*Type-of-Death*) level requires features with greater discriminative abilities compared to the coarse-grained (*Time-of-Death*) level categories. Considering the inherent problems that characterise the VA corpus as already discussed and summarised in table 6.2, this may have resulted in more overlapping and less discriminating features, which tended to pose problems for the Machine Learning algorithm at the fine-grained level.

We also acknowledge that the procedure regarding *locally-semi-automatic* approach to feature reduction in Chapter Seven and the collocation extraction based features should have been done on a separate training set for each of the 10-fold cross validation and not the whole corpus, to ensure terms extracted were mutually exclusive. We wish to state this as a limitation even though the difference in results may not be significant.

8.7 Summary

The objective of the work set out in this Chapter was to improve on the results obtained in Chapter Seven. Chapter Seven established the feasibility of predicting the cause of death from the open narrative part of VAs. This Chapter described experiments which that explored the feature space of the open narrative text, which resulted in significant improvements in performance over what was obtained in the previous Chapter.

A detailed analysis and discussions were done on the behaviours and the effects of the various features that were employed in the development of the model. It was also found that the performance accuracy at the *Time-of-Death* level was relatively higher compared to the *Type-of-Death* level and the reasons that may have accounted for said difference were also discussed.

Areas that required exploration in order to improve on the current results were also discussed.

Chapter 9

The integration of domain knowledge with linguistic and lexical-based features for Cause of Death prediction

"Knowledge is of two kinds. We know a subject ourselves, or we know where we can find information upon it", Samuel Johnson

9.1 Introduction

In Chapter Eight we explored various features derived from the contextual information available in the open narrative text of VA documents. This contextual information was obtained through statistical and linguistics based approaches to analysis of the content of VA open narrative text. In other words, no specific domain knowledge was employed in obtaining these features.

In this Chapter we consider information derived from the closed responses part of the VA. We argue that the closed responses part contains specific domain knowledge derived from physicians and VA experts. The rationale for this argument is that the questions are carefully designed to elicit information specific to the causes of death based on the knowledge physicians have about the diseases that result in the death. This includes associated signs and symptoms that may have led to the death of the individual.

However, as explained in Chapter One, the closed questions are invariably limited in capturing all relevant information and existing automatic methods have also ignored the information available from the open narrative part. This Chapter explores the extent to which the integration of knowledge obtained from the both the closed response and the open narrative impacts on performance accuracy of our automated approach. To achieve this, two set experiments are described. The first set of experiments considered only the closed response part of the VA data which is referred to as VAModel2. The second set of experiments were based on the combination of the open narrative and closed response parts which is referred to as VAModel3. The experiments described in this Chapter are based on the model from Chapter Eight (SVM algorithm and Groupings3) due to it having been established as the best performing model.

9.2 Pre-processing and Feature Value Representation of Closed Response.

It was demonstrated in Chapter Seven that there is a strong impact of feature value representation on the performance of Machine Learning algorithms. From this background, we began the experiments with investigation into the suitable feature value representation scheme for the closed responses part of VAs. We investigated two approaches in this experiment: *dichotomised* and *nominal* approaches. The *dichotomised* was one employed by Byass et al., (2006), where variables (features) were dichotomised as part of the pre-processing step of classifier development. The *dichotomised* approach to representation tended to capture only positive responses to symptoms and ignore other response options captured. In contrast, the *nominal* approach took into account all information related to the options provided and collected them as part of the closed responses for training algorithms. The inclusion of all response options available was based on the hypothesis that those variables could contain additional information that may be relevant to a Machine Learning algorithm. We illustrated this with an example, as shown in Figure 9.2. All information relating to demographics was removed. Any other information was considered in building the model, which included the information (variables), recommended by WHO as the most relevant variables and used by some existing approaches (Byass et al., 2006).

Was the baby born alive or dead	1. alive	2. dead	8.don't know

Figure 9.1 options for closed response part of Verbal Autopsy

Figure 9.1 shows a sample question contained in the closed response part of the VA questionnaire that sought to elicit information from a woman as to whether her child was born alive or stillborn with the various response options available. Our proposed approach adopted in this experiment represented the variables as captured on the VA questionnaire.

```
@ATTRIBUTE        BORNALIVE          {1,2,8}
```

Figure 9.2 WEKA arff file format of nominal feature representation

@ATTRIBUTE	BORNALIVE	{1 }

Figure 9.3 WEKA arff file format of dichotomised feature representation.

As Figure 9.2 demonstrates where all the options were considered in the feature value representation, the *dichotomised* approach considered only the *alive* option for the representation scheme with feature value *1* as shown in Figure 9.3. The other two options, *dead* and *don't know,* are ignored. Two datasets were created based on these two approaches, which served as inputs to our SVM-based algorithm.

9.3 Combination of Closed Response and Open Narratives Text

This experiment aimed at exploring the extent to which open narrative text contributed to the overall performance accuracy of predicting cause of death from VA. Having established the effective feature representation scheme and the best performing feature set for the closed responses, we created a final dataset that combined the features extracted from the open narrative and the closed responses. The parameter settings and evaluation methods employed in conducting the experiments as described in Chapter Seven remained unchanged.

9.4 Results

9.4.1 Performance Comparison: Dichotomised vs. Nominal Feature Value Representation for Closed Response

Table 9.1a Time-of-Death results: performance comparison-dichotomised vs. nominal representation schemes for closed responses

representation scheme	percentage accuracy	macro-average -f-measure
dichotomised	97	0.97
nominal	96	0.96

Table 9.1b Type-of-Death results: performance comparison-dichotomised vs. nominal representation schemes for closed responses

representation scheme	percentage accuracy	macro-average -f-measure
dichotomised	67	0.65
nominal	77	0.75

Tables 9.1a and 9.1b show the results obtained from an experiment that sought to compare the performance between the *dichotomised* and *nominal* representation schemes, applied to *Groupings3*. As the tables show, the performance for the *dichotomised* scheme was significantly better than the *nominal* approach for *Time-of-Death* (accuracy: 1 % (97 - 96); macro-average-f-measure: 0.1 (0.00 - 0.96); p-value >0.001). In contrast, *nominal* performed significantly better than the *dichotomised* approach for *Type-of-Death* (accuracy: 10% (67 - 77); macro-average-f-measure: 0.10 (0.65 - 0.75); p<0.001).

9.4.2 Performance Comparison: Open Narrative Text vs. Closed Response

In this section we compare the performance achieved between the open narrative and closed response parts for both *dichotomised* and *nominal* representation schemes. We repeat the best results obtained for open narrative here to facilitate comparison.

Table 9.2a Time-of-Death results: performance comparison- open narrative vs. closed response

model	percentage accuracy	macro-average-f-measure
open narrative	85.4	0.85
closed response - nominal	96	0.96
closed response - dichotomised	97	0.97

Table 9.2b Type-of-Death results: performance comparison- open narrative vs. closed response

model	percentage accuracy	macro-average-f-measure
open narrative	46.3	0.39
closed response - nominal	77	0.75
closed response -dichotomised	67	0.65

As tables 9.2a and 9.2b show, models based on the closed response part (both dichotomised and nominal representation schemes) performed significantly better than the open narrative for both *Time-of-Death and Type-of-Death.* For the open narrative text and closed response part under the *dichotomised* representation scheme: *Time-of-Death (*accuracy: 11.6% (97 - 85.4); 0.12 (0.97-0.85); p<0.001) and *Type-of-Death* (accuracy: 20.7% (67 - 46.3); 0.26 (0.65 -0.39); p<0.001). Similar performance pattern was observed for the *nominal* representation scheme: *Time-of-Death* (accuracy: 10.6% (96 - 85.4); 0.11 (0.96 -0.85); p<0.001) and *Type-of-Death* (accuracy: 30.7% (77 - 46.3); 0.36 (0.75 - 0.39); p<0.001).

9.4.3 Performance of Combined Model

In this section we examine the combining of information obtained from the two sources: open narrative text and closed response part, referred to as VAModel3, the *combined* model. Our exploration considered the *nominal* representation for the closed response part because even though the *dichotomised* performance was better than *nominal* representation by 1% (p=0.002), *nominal* representation was demonstrated to be equally robust for *Time-of-Death* and even outperformed *dichotomised* for the *Type-of-Death* categories (p-value<0.001).

Table 9.3a Time-of-Death results: combined model

model	percentage accuracy	macro-average-f-measure
closed response - nominal	96	0.96
combined	96	0.96

Table 9.3b Type-of-Death results: combined model

model	percentage accuracy	macro-average-f-measure
closed response - nominal	77	0.75
combined	78	0.76

Tables 9.3a and 9.3b show the performance obtained for the combined model for *Time-of-Death* and *Type-of-Death* respectively. It was observed that no change in performance was seen for *Time-of-Death* when compared against the closed response. However, although not statistically significant, a slight increase in performance was observed for *Type-of-Death* (accuracy: 1% (78 - 77); 0.1 (0.76 - 0.75); p=0.09). We therefore examined this in detail at the category levels to gain insight into the results obtained.

Table 9.4a Time-of-Death results: performance comparison - closed vs. combined by category

Time-of-Death	percentage accuracy		macro-average-f-measure	
	closed	combined	closed	combined
Neonatal	93	93	0.94	0.94
PostNeonatal	94	95	0.93	0.93
Stillbirth	99	99	0.99	0.99

Table 9.4b Type-of-Death results: performance comparison closed vs. combined by category

Type-of-Death	percentage accuracy		macro-average-f-measure	
	closed	combined	closed	combined
Neonatal_infection	68	71	0.69	0.70
Prematurity	84	83	0.83	0.82
Antepartum_stillbirth	93	92	0.92	0.92
Intrapartum_stillbirth	86	87	0.87	0.88
Pneumonia	61	64	0.60	0.63
Diarrhoea	68	65	0.60	0.58
PostNeonatal_other_infections	56	54	0.54	0.54
Birth_asphyxia	82	82	0.79	0.79
Neonatal_other_causes	37	36	0.43	0.43
Malaria	47	47	0.49	0.50

Table 9.4a and 9.4b show the breakdown performance by classes of both *Time-of-Death* and *Type-of-Death* respectively for the combined model. The performance values obtained for the breakdown for *Time-of-Death* confirmed the results obtained in table 9.3a. In contrast, three different performance trends were observed for the *Type-of-Death* categories. Firstly, it was observed that there was an increase in performance for the *combined* model when compared with the closed response part only for: *Neonatal_infection* (accuracy: 3% (71 - 68); 0.1 (0.70 - 0.69); p=0.15); *Intrapartum_stillbirth* (accuracy: 1% (87 - 86); 0.1 (0.88- 0.87); p=0.26) and *Pneumonia* (accuracy: 3% *(*64 - 61); 0.3 (0.63- 0.60); p=0.37). This suggested that open narrative text tended to improve performance for those cause of death classes. Secondly, no change in performance was observed between closed response and combined models for *Malaria* and *Birth_asphyxia.* This also suggested that there was no additional useful information obtained from the open narrative text for those cause of death classes. Finally, there was a reduction in performance accuracy of the combined model for other cause of death classes: *Prematurity* (accuracy: -1% *(*83 - 84); -0.1 (0.82 - 0.82); p-value=0.07); *Antepartum_stillbirth* (accuracy: -1% *(*92 - 93); 0 (0.92-0.92); p=0.02); *PostNeonatal_other_infections* (accuracy: -2% (54 - 56); 0.1(0.61 - 0.62); p=0.01); *Diarrhoea* (accuracy: -3% (65 - 68); -0.2(0.58 - 0.60); p<0.001); *Neonatal_other_causes* (accuracy: -1% (36 - 37); 0(0.43 - 0.43); p=0.13).This suggested the open narrative text did not add any useful information to the closed part but rather degraded the performance of the *combined* model.

9.4.4 Error Analysis

Our analysis has so far centred on the performance accuracy of the classification models. It is however imperative to examine the rate of misclassification in order to gain a better understanding of the performances of the models. As discussed in Chapter Four, *false positive rates* (FPR) allow us to perform an error analysis to determine the rate of misclassifications (Fawcett, 2006). A false positive suggests a prediction of cause of death that is not actually the cause. Thus, a good model is required to obtain a relatively low FPR.

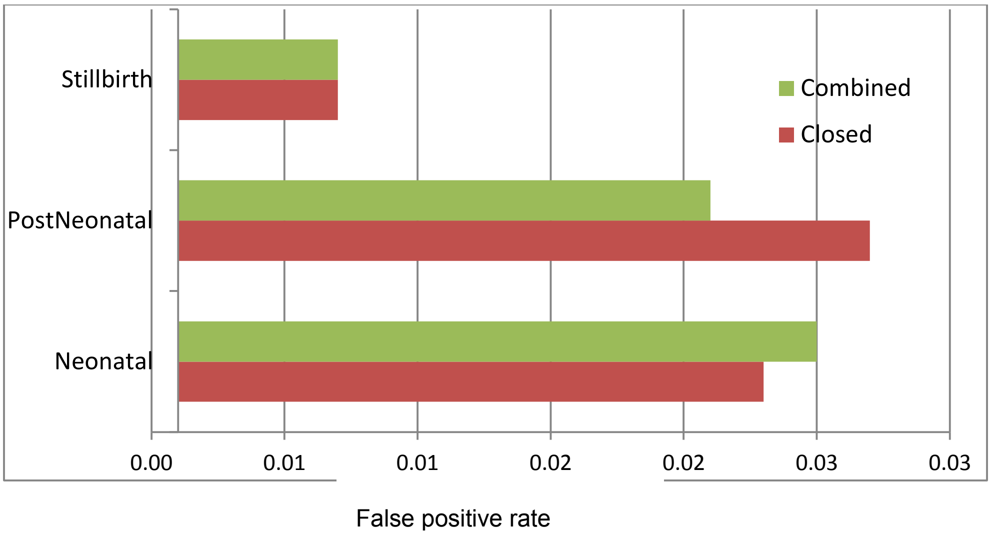


Figure 9.4a Time-of-Death results: a comparison of FPR

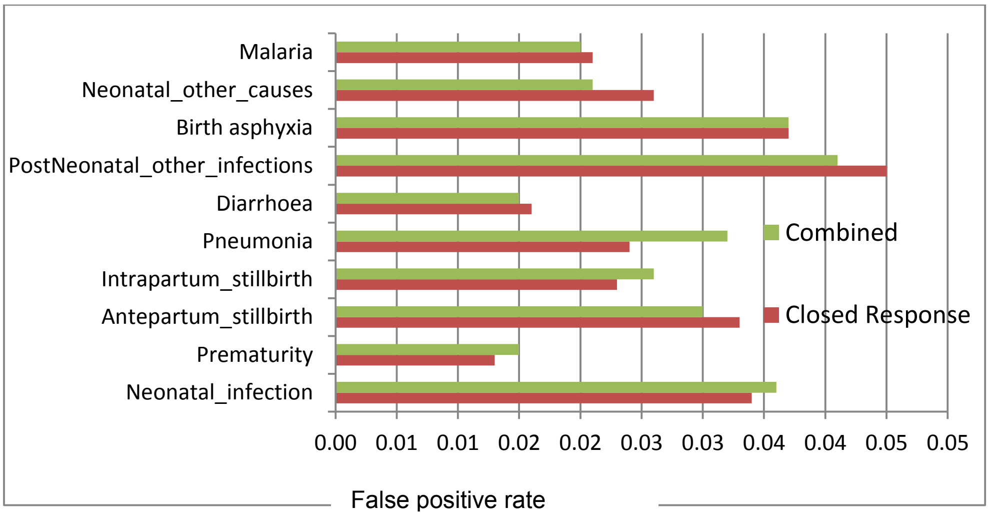


Figure 9.4b Type-of-Death results: a comparison of FPR

Figures 9.4a and 9.4b show FPR of both combined and closed part models for *Time-of-Death* and *Type-of-Death* respectively. With regard to closed and combined models for *Time-of-Death*, it was observed that the misclassification rate for the closed response and combined model remained unchanged for *Stillbirth* (0.01). However, with regards to *PostNeonatal* category, there was a slight reduction in FPR for the combined (0.02) compared with closed (0.03), suggesting open narrative helped in avoiding some misclassification of *PostNeonatal* deaths. On the contrary, for the *Neonatal* category, there was a slight increase in FPR for the combined (0.03) compared to closed response (0.02), suggesting that a combination of open narrative and closed responses could result in some misclassifications.

Similar patterns were observed in the *Type-of-Death* categories as Figure 9.6b demonstrates. The combined model obtained slightly less FPR when compared to the closed, suggesting some benefit of open narratives in avoiding misclassifications for these classes: *Malaria* (combined (0.020); closed (0.021)); *Neonatal_other_causes* (combined (0.020); closed (0.026)); *PostNeonatal_other_infections* (combined (0.041); closed (0.045)) and Diarrhoea (combined (0.015); closed (0.016)), *Antepartum_stillbirth* (combined (0.033); closed (0.030)). However, there were some marginal increases in the misclassification rate of the combined compared to the closed response model, suggesting that open narrative tends to cause some misclassification for some other category: Pneumonia (combined (0.032), closed (0.024)), *Intrapartum_stillbirth* (combined (0.024), closed (0.032)) and *Neonatal_infection* (combined (0.036), closed (0.034)). No misclassification was observed between combined and closed for *Birtt_asphyxia* category.

9.5.5 Model Reliability Test

The 10-fold cross validation approach used in this research raises questions about the reliability of the models, considering the characteristics associated with the k-fold cross validation methods as discussed in the Chapter Four. In view of this argument, it is imperative to know the reliability of the models on different datasets and settings. To determine the reliability of our models we employed the kappa statistic to measure whether the level of agreement between the models and the gold standard was due to chance (Fielding and Bell, 1997). Table 9.5 gives a summary of the kappa based on *Groupings3* results.

Table 9.5a Time-of-Death comparison of kappa statistic values

Model	Kappa statistic
Baseline	0.0
Open narrative	0.53
Closed response	0.94
Combined	0.94

Table 9.5b Type-of-Death comparison of kappa statistic values

Model	Kappa statistic
Baseline	0.0
Open narrative	0.4
Closed response	0.71
Combined	0.75

Tables 9.5a and 9.5b show the kappa statistic values obtained for the various models. As the values suggest, in terms of *Time-of-Death*, the level of agreement was the same for both closed and combined models, each obtaining 0.94 whilst the open-narrative-based model also obtained a kappa score of 0.53. With regards to *Type-of-Death*, the combined model obtained a kappa score of 0.75 whereas the closed response model obtained 0.71. The open narrative also demonstrated a moderate agreement (0.4), which is better than the majority baseline (0.0). These values are discussed in section 9.6.4.

9.6 Discussion

9.6.1 Performance Difference: Closed Response Feature Value Representation

The results obtained from the experiments tend to confirm our rationale for the need to accurately represent the responses to the closed questions for VA. Based on the results, it could be argued that the *dichotomised* representation scheme has the potential of missing information required by the Machine Learning algorithm, especially at the fine-grained level where detailed type of death information is critically needed. This scheme presented only positive responses. For example, Byass et al.(2006) method required that duration of illness must be calculated based on the definitions and cut-offs for *acute* (less than two weeks) and *chronic* (more than two weeks) for fever. In contrast, the nominal representation scheme was demonstrated to contain more information relevant to the Machine Learning algorithm. This enabled true patterns of response to be generated and learned by the algorithm. It is our view that this is an important and novel research finding for all automatic approaches to predicting cause of death from VAs.

9.6.2 Performance comparison: open narrative vs. closed response

The relative higher performance accuracy obtained by the closed response part of the VA seems natural due to the following reasons:

- As indicated, the closed response questions were carefully designed by the domain experts to elicit information about signs and symptoms that were specific to a given cause of death. The responses to the targeted questions were indicative of the diseases that caused the death.

- The level of noise for the closed responses could be relatively low compared to the open history. The closed response part had fewer options to select from, making it relatively less prone to the introduction of noise into the feature space. It must however be pointed out that the limitation of closed response remains; the inability to capture all relevant information as we have partially demonstrated from the results obtained.

Furthermore, the results obtained from the open narrative demonstrated the potential that exists in the use of the narratives as an alternative in predicting causes of death. The current results are characterised by several factors which may have affected the performance of the Machine Learning algorithm as discussed in Chapter Five of this thesis. Some of the striking problems may have contributed significantly to the performance accuracy and these include:

- There was a lack of standardisation of the language used in the reporting of VA. This can be put into context when compared with the language used in other biomedical settings. The language used in the biomedical setting was constrained by the limited vocabulary of that domain. This was however not the case in the VA domain as discussed in Chapter Six and Seven.

- The large proportion of contextual spelling errors observed in the VA open narrative could also account for the relatively low performance. Even though our transcription software had inbuilt spell-checker which was adapted from Microsoft Office word processing software, this was found to be ineffective upon visual inspection of a sample text as it was heavily dependent on the judgement of the user for correction which was problematic considering the large scale and the volume of the transcription. For example a user could mistakenly make a wrong choice from the list of words provided by the transcription software. As a further check we employed the noisy channel spelling correction algorithm to automatically correct any errors, but this resulted in lower performance accuracy. This could have been due to high false positive corrections that were automatically made by the algorithm. For example, spell correction of a wrongly spelt "vergina" resulted in "virginia" as opposed to "vagina", which in turn had an adverse impact on the algorithm. Further pre-processing such as carrying out a much more complex spell correction might improve the current results. Also contextual spelling errors could not be detected.

Again, it could be argued that the open narrative corpus currently lacks additional knowledge apart from the domain knowledge derived from the cause of death annotation by the physicians. There is an additional need for annotations which could be obtained from the corpus through semantic analysis and further annotations of concepts in the text. For example the GENIA corpus had various annotations: Part-of-Speech; co-references; semantic classes and syntactic information (Tateisi et al., 2005). These annotations tended to make the GENIA corpus more useful, which in turn led to an improved performance accuracy of GENIA corpus trained models. Similar schemes could be adopted to enrich the VA corpus with the aim of improving the performance accuracy of VA trained models.

Nevertheless, with these limitations and in the absence of additional information, it can be argued that the current performance obtained by the open narrative model suggests a viable alternative or complementary approach to the closed response approach to predicting cause of death.

9.6.3 The Benefit of Domain Knowledge from Cause-of-Death Re-Groupings

The results obtained from the re-groupings further underscored the need for more domain knowledge in model development in order to improve on the performance accuracy. It is therefore not surprising that among the three groupings employed in our experiments, *Groupings3*, which was created with inputs from domain experts at the LSHTM, turned out to be the best performing one among the three groupings. This grouping contained additional information which further removed some of the category noise that may have been introduced during annotation and re-groupings, which could have caused some side-effects for the learning algorithm as discussed in Chapter Four.

9.6.4 Prediction Errors and Model Reliability

The lack of standardised language and the vague expression of medical concepts, as discussed in Chapter Six, could account for the relatively higher misclassification rate observed for the open narrative. These expressions lacked precision which resulted in overlapping concepts. As discussed in Chapter Five, these overlapping features were potential sources of noise, and this could have been problematic for the learning algorithm when it was attempting to clearly determine the decision boundaries for the various causes of death within the feature space. Consequently, this resulted in the variance phenomena discussed in Chapter Four. One possible solution to this problem would be to acquire more training data as pointed out by Friedman (1997).

Furthermore, the relatively low misclassification rate observed for the closed response part and the combined models suggests carefully tailored questions, which are guided by the domain knowledge of medical experts. However, the relatively low misclassification rate observed for some categories of the combined model compared to the closed response model tends to suggest that open narrative can be useful in avoiding a high rate of misclassification. Intuitively, this makes sense and is somewhat expected since the open narrative should contain contextual information helpful to the classifier. On the contrary, in cases where the relatively high misclassification rate was observed for combined compared to the closed, additional pre-processing and feature engineering would be needed. For example, carrying out a feature reduction as a prior step to training the classifier would be a useful step. It would therefore be interesting to examine this in detail using other Computational Linguistics methods to explore this further.

With regards to the reliability of the models, Hripcsak and Heitjan (2002) suggested that kappa values lay between 0 and 1, where 0 was an indication of agreement by chance and 1 suggested a perfect agreement. However, negative values were also obtainable, and they indicated disagreement (Hripcsak and Heitjan, 2002). Moreover, the reliability of the model as determined by the kappa statistic values obtained could be interpreted and contextualised using the grading system proposed by Landis and Koch (1977) for the biomedical domain. The values obtained by the models suggested that all predictions were not by chance as per the grading system; the open narrative obtained a moderate agreement (0.4) whereas the closed response and the combine models value suggested a significant to excellent agreement (0.71) and (0.75) respectively.

Notwithstanding the kappa values obtained, Hripcsak and Heitjan (2002) cautioned that these values should be interpreted with a careful consideration of the domain and the context. For example, the average level of agreement between physicians obtained for this dataset, which is the same dataset used for this research, was 0.64 (0.1 - 0.9) as reported by Edmond et al.(2008). Also, the level of agreement observed using the interVA methods (Byass et al., 2006) to predict causes of death using the closed responses part was only kappa = 0.4. This validation was done during a workshop to review and analyse infant death using physicians and InterVA version four, organised by the World Health Organisation in Geneva. Considering the available results obtained from different approaches on the same dataset, it is evident that the results obtained are contextually reasonable and comparable to other studies within the VA domain, where low agreements tend to be reported (James et al., 2011). However, a direct comparison between our results obtained from our research and other studies is not a straightforward process due to the differences that exist between the causes of death categories. For example our results contained a fined-grained stillbirth sub-category, which was usually not reported or was combined and reported as stillbirth by other automated approaches.

9.7 Summary

This Chapter explored the extent to which integration of knowledge derived from domain experts such as physicians would help contribute to improving the performance accuracy of our models. From the model developed in Chapter Eight, which was based on only open narrative text of VA, two additional models were developed: one was based on only closed response and the other based on a combination of both open narrative and closed response. It was argued that the closed response model was considered to contain domain specific knowledge whereas the model based on the open narratives contained contextual information obtained from the information provided by the respondents. The third model, which was based on the combination of feature sets from the two models enabled us to observe the differences in performance in order to assess the impact of the information obtained from the open narrative response on performance and vice versa.

We proposed a nominal feature value representation scheme for the closed response part as an alternative to the dichotomised approach employed by the existing automatic approaches to VA analysis discussed in Chapter Two. It was observed that our proposed scheme outperformed the dichotomised representation at the *Type-of-Death* level with competitive performances at the *Time-of-Death* level. The reasons that account for this were discussed.

Additionally, various performance measures were explored to determine the performance of the models. It was shown that the model based on closed response on average outperformed the model based on the open narrative, with a marginal increase in performance for the combined model. However, a detailed analysis carried out also revealed the benefit of integrating information from open narrative and closed responses. For example, a combination of closed and open narrative was demonstrated to be beneficial for *Postneonatal* deaths at *Time-of-Death* level and *Neonatal infections* and *Intrapartum stillbirth* at the *Type-of-Death* level.

Finally, it was also demonstrated that cause of death category labels developed based on domain knowledge tended to remove some noise that could exist, which resulted in further improvements in performance.

Chapter 10

Conclusions and Future Work

"If we knew what it was we were doing, it would not be called research, would it?"

Albert Einstein

10.1 Introduction

Automatic prediction of cause of death from VA remains an active research area due to its potential benefits to healthcare in general and developing countries in particular, where VAs are predominately employed. We begin this Chapter by providing a summary of the issues discussed in the various Chapters of this thesis.

Furthermore, the aims, objectives and main contributions of the PhD research as outlined in Chapter one are assessed to demonstrate how these were achieved. This Chapter finally draws conclusions from the experiments carried out and provides suggestions on some potential future work that could be carried out to improve on the results obtained from this research.

10.2 Summaries of Chapters

10.2.1 Chapter 1

In Chapter One, we set the scene for the research project discussed in this thesis. A brief introduction to the domain of VAs was given, which led to the discussion in relation to the rationale and the motivation for the research project. The aims and objectives for the project were outlined, as well as the novelty and the contributions of this thesis to various research groups and also the impact on VA research.

10.2.2 Chapter 2

This Chapter provided a survey of the VA domain; an overview was given, which described the data collection processes and the various uses of VA information. A comprehensive survey of the literature was also carried out, which focused on the existing automatic approaches and how these approaches differ from our approach, which is formulated as a Text Classification problem.

10.2.3 Chapter 3

In Chapter Three we explored the various approaches and processes to Text Classification. We particularly focused on the feature engineering phase of the process and identified proven methods and techniques available from the literature for extracting and representing features for classification of textual data. This Chapter concluded by conducting a survey of studies that had applied these techniques within the biomedical domain and text originating from different settings within that domain.

10.2.4 Chapter 4

In Chapter Four we explored the field of Machine Learning and the various Machine Learning algorithms employed in carrying out classification tasks. We further surveyed the literature to identify various evaluation metrics that were employed in determining the performance of a given Machine Learning algorithms and their overlaps in performance evaluation. Finally, the various issues that tend to adversely impact on the performance of Machine Learning algorithms were discussed.

10.2.5 Chapter 5

This Chapter was concerned with the surveillance system that was established as part of a large scale field trail which ran over a nine year period to facilitate the identification of the deaths which enabled the VA corpus to be built. A detailed description of the methods employed to generate and manage the VA was provided in this Chapter. We further examined the original causes of deaths that were obtained as part of the corpus and proposed three groupings based on the classification scheme originally employed and also advice from domain experts. These groupings formed the basis of our experiments.

10.2.6 Chapter 6

In Chapter Six, we carried out an analysis of the VA corpus: a detailed description of the language was given and the issues and their implications for computational research were discussed. The issues identified in this Chapter informed the methodology employed in carrying out the research and these are subsequently summarised.

10.2.7 Chapter 7

In this Chapter we reported the initial work that was carried out to explore the problem space of VAs. The objectives were to determine the suitability of the various Machine Learning techniques identified in the course of the literature survey as previously discussed. The experimental results suggested the SMO, which is a refinement of the Support Vector Machine, as the most suitable Machine Learning algorithm for this VA domain among other popular algorithms such as the *Naive Bayes* and *Random Forest*. Various feature value representation schemes were also explored: *Binary*, *Term Frequency*, *Normalised Term Frequency (NTF)* and *TFiD*. The experimental results suggested *NTF* as the best performing representation scheme for the VA domain. The Chapter further described another experiment which aimed to identify the most predictive words for each cause of death category based on the log-likelihood statistical measure, referred to as *locally-semi-automatic method* feature reduction strategy. This was compared with *String-matching* and *Information Gained*; the *locally-semi-automatic* approach was demonstrated to have achieved better performance.

10.2.8 Chapter 8

This Chapter discussed experiments conducted to further improve on the performance accuracy obtained from the previous experiments. The model developed from this set of experiments was referred to as VAModel1. The motivation and the choice of features were discussed. Furthermore, a detailed analysis of the behaviour and impact of the various features on the performance were conducted. It was observed that the *simplified relative word position* feature, which potentially captures the sequential order of words in the text and phrases extracted through collocation bigram based on ranked log-likelihood statistics (Dunning, 1993), was relatively the best performing feature. This was based on the top-3 ranked words from the collocate list, which also contributed significantly to the overall performance. The less performing features such as noun and verb phrase patterns were also observed and the reasons for their underperformance were discussed. In summary, this Chapter achieved the first aim set out in this thesis which was to establish the feasibility of predicting cause of death from the open narrative text of VA.

10.2.9 Chapter 9

This Chapter reported on experiments conducted to establish the second aim set out in this research project, which was to assess the extent to which the open narrative could contribute to the performance accuracy of predicting causes of death from VA. To be able to carry out this assessment, there was the need to determine the performance accuracy obtainable using only the closed responses part of the VA data, called VAModel2. However, prior to this, initial experiments was conducted to determine the suitable feature value representation scheme (nominal) for the closed responses of which our proposed scheme outperformed the existing representation schemes (dichotomised). Our proposed scheme allowed all valid values to be represented as nominal, which differed from schemes employed by almost all the existing automatic approaches, where the responses were dichotomised (Flaxman et al., 2011; Byass et al, 2010). This Chapter further presented and discussed the performance difference observed between the models based on open narrative, closed response and a combination of the two sources called VAModel3. It was further established from the results that domain specific knowledge contributed to an increase in the performance of a Machine Learning algorithm, for example by carefully selecting features or variables based on advice from VA experts. Subsequently, it was further argued that the closed questions represented some form of domain knowledge since these were developed by physicians with experience in VA research, and this resulted in the closed response achieving a relative better performance over the open narrative.

10.3 Research Aims and Objectives: Stock Taking

Chapter One outlined the main research aims and objectives we intended to achieve in this research as follows:

10.3.1 Aims

We aimed to establish the feasibility of automatically predicting cause of death from VAs open narrative text. The cause of death was categorised into two levels. *Time-of-Death* was concerned with the period in which the death occurred. This level of death information is crucial for health planners and researchers. The second level was concerned with the feasibility of prediction of *Type-of-Death* categories, which contain additional information about the type of disease or illness that led to the death. This aim was achieved through a series of experiments described in Chapter Seven and Eight of this thesis. An overall performance accuracy between baseline and achievement for *Time-of-Death* (accuracy: 43.1% (85.4 - 42.3); macro-average-f-measure: 0.29 (0.25 - 0.54); p=0.01) and *Type-of-Death* (accuracy: 21.8% (46.3 - 24.5); macro-average-f-measure: 0.31(0.39 - 0.08) p =0.007) was recorded. Additionally, a reliability test using kappa statistic also suggested a value of approximately 0.53 and 0.4 for *Time-of-Death* and *Type-of-Death* respectively. These values suggested that the performance accuracies obtained were not by chance (Hripcsak and Heitjan, 2002).

The second aim of this research was to determine the extent to which information obtained from the open narrative text part of VA could improve the accuracy of automatic prediction of causes of death. In Chapter Nine, we described experiments carried out to achieve this aim. The results obtained suggest that information contained in the open narrative text has the potential to improve prediction accuracy for some causes of death categories whereas information contained in the closed-response part is enough for other categories. For example, it was observed from the results that a combination of the sources of information led to an improvement in accuracy for *Neonatal infections* and *Intrapartum stillbirth*. However, it was pointed out that these results were inconclusive as there was more room for research and exploration of the open narrative text space, something which is discussed later in this Chapter.

10.3.2 Objectives

One of the objectives set out to be achieved in this research was to build a corpus of VA text to serve as a resource for research. In Chapter Six, a detailed description was given regarding how a corpus of 2.5 million words had been built. This corpus could be used in other language research to further explore areas of interest once ethical issues related to its use have been cleared.

Another objective that has been achieved is the findings regarding the Machine Learning algorithms and techniques suitable for the VA domain. This is the only research that has extensively investigated the various factors that impact on the performance of a Machine Learning based prediction model: algorithms; feature value representation and feature reduction strategy. For example, even though Machine Learning algorithms such as the Random Forest have previously been explored (Flaxman et al., 2011), this research proposed an SVM based method, which has been demonstrated to be superior to Random Forest and a viable Machine Learning algorithm for the VA domain. The experiments and the findings were subsequently discussed.

We also set out to investigate the features and methods for extracting suitable features in order to predict causes of death from VAs open narrative text. Considering the fact that over 40 countries around the world use VAs with different language and settings, one of the motivations was to investigate features that could be replicated in other languages and settings. The objective was therefore to identify features that would be language independent. In Chapter Eight, we described the statistical and linguistic based approaches that were employed to identify and extract features used in building our prediction model. These features could be considered language independent and theoretically, should be applicable in different language environment and settings. However, this could not be tested and thus could be explored in future work.

Another objective that was set out was to disseminate the findings through various forums: journal publications, conference proceedings, workshops and seminars. Presentations were made at six conferences to both Corpus and Computational Linguistics audiences and researchers as well as VA researchers and practitioners.

10.4 Implications for Current Practices

The open narrative text has been an integral part of the information collected for VA and has been used in determining the cause of death by physicians. However, as indicated in Chapter Two recent development of automatic approaches has so far been limited to the closed response part of the VA information. Consequently, there is a growing interest in how the information contained from the open narrative could be exploited using automated approaches. The results obtained from this research have demonstrated the feasibility of automatically predicting from the open narrative of VA. This offers an alternative to predicting cause of death using the closed response only. Also, similar to the physician approach, this research has demonstrated the feasibility of using the open narrative as a complement to the closed response for automated approaches.

However, the potential implementation of this research is dependent upon the quality of the open narrative text. This implies the need for the establishment of a robust data collection system that will ensure the open narrative text that is generated is of good quality. For example, lack of standardised language used in reporting VAs was found to be critical for our Machine Learning approaches. This requires the development of some glossary of terms and training of data collectors to use. The current collection process for the open narrative could be improved, and in turn could improve on the quality of the text and efficiency.

It must be pointed out that efforts are already being made by VA researchers in this regard. For example, at the 2nd Global Congress on VA (14 – 16 October 2014, Rhodes, Greece)[1,] where we also presented a talk titled “Machine Learning approaches to predicting cause of death from verbal autopsy open narrative text: the vector method”, a recent pilot study was also presented by other groups of researchers, which aimed at exploring the possibility of reducing the amount of text an interviewer would write during a VA interview. They proposed a checklist of relevant symptoms which would be checked by the interviewer whenever a respondent mentioned any of those symptoms found on the check list during narration. This approach could cut down on the amount of text and eventually reduce the amount of errors in text.

However, the proposed approach poses practical challenges for the interviewers such as cognitive pressures in processing and recording the information at the same time without losing some of it in the process. It also requires a relatively higher level of education and skills set to be able to function effectively, which is mostly unavailable in the settings where VA is

---

[1]http://globalva.org/conference-agenda

employed. It is therefore imperative to explore other cost effective and efficient methods of collecting the open narrative without compromising on its quality. For example, the method of a data collector manually writing the interview on the form before it is transcribed onto a computer for processing could be eliminated by taking advantage of voice-to-text technology. Other technological options could be explored in this regard.

10.5 Limitations and Future Work

10.5.1 Corpus Sample Size and Generalizability

The results obtained from the experiments are based only on the infant corpus, which comprises of stillborn and children up to one year of age. Also, as indicated, this corpus is a sample collected from only seven districts in Ghana. Even though our approach aimed at developing a model that would be generalizable in different contexts and settings, something which the results obtained and the kappa statistics values suggest is true, we do acknowledge that this remains a limitation until further experiments can be carried out to determine the generalizability of our approach.

Also, as indicated, in the absence of a large training set, our experiments employed the stratified 10-fold cross validation approach to evaluate the performance of our models. Although 10-fold cross validation has been established as a valid approach that tends to produce results comparable to results obtained from unseen data (Kohavi, 1995), we wish to state this as a limitation until an independent test data set is used for further testing.

Furthermore, it was noted in Chapter Three that over 40 countries employ VA as an approach to determining cause of death. There is therefore an opportunity for further validation studies to assess the performance within multinational and multi-cultural settings. This would help to improve on the current performance and the generalizability, which would result in obtaining a model that is compatible with varying contexts.

10.5.2 Further Annotations of Corpus

Our current VA corpus is annotated at the document level, which limits the potential usefulness of the corpus. For example, in Chapter Six we pointed out that the GENIA corpus has been employed in carrying out numerous studies outside the purpose for which it was originally built due to the various levels of annotation that have been incorporated (Kim et al., 2003). For example,Tsuruoka et al. (2005) aimed at improving the usefulness of the GENIA corpus by incorporating annotations of syntactic structures into a subset of the corpus. The potential exists to establish further types of annotation over the VA corpus for further annotations in order to improve its usefulness for NLP research.

10.5.3 Opportunities for Further NLP and Machine Learning Research

In Chapters Seven and Eight, we identified some of the opportunities for further research in the area of NLP and Machine Learning. We demonstrated that the current state-of-the-art PoS taggers could not achieve the 97% accuracy, which is the performance required as a component of an NLP system. Our results suggested the current best performing PoS tagger achieved 83 % accuracy, which demonstrated a clear need for adaptation and an evaluation study into how these taggers could be improved on the VA text. For example, Tsuruoka et al.(2005) again demonstrated the need to adapt a general purpose PoS tagger to be able to achieve the required performance accuracy for biomedical text. Similar approaches could be explored and a PoS tagger developed for VA text.

Furthermore, this research employed a supervised Machine Learning approach to predicting causes of death from VA. However, other Machine Learning paradigms exist which could be explored and results compared. For example, considering the expensive and time consuming task of acquiring a gold standard dataset, unsupervised Machine Learning and semi-supervised methods could be explored as an alternative (Gentleman and Carey, 2008) and the results compared to those obtained in this research. Also, cost sensitive learning could be explored to ensure rare categories that are not equally represented by Machine Learning algorithms are not ignored, as discussed in Chapter Four.

10.5.4 Further Exploration of Feature Space

As has been pointed out throughout this thesis, this research established the feasibility of predicting cause of death from the open narrative text of VAs. It is therefore imperative to note that more work is required to fully exploit the information available in the open narratives. This would require further feature engineering approaches to the exploration of the feature space of the open narratives. For example, the features explored in our experiments have been limited to only information derived from the VA text. Therefore, information could be obtained from external sources such as specialised ontology for VAs, which could be developed and used to further improve on the performance. The use of information obtained from external sources to enhance the performance of text classification models has been proven to be a useful approach (Gabrilovich and Markovitch, 2006). However, considering the fact that VA text is obtained from informal settings makes it problematic to employ existing resources such as ontologies as part of its model development. This problem has been recognised and been addressed by Smith and Fellbaum (2004). They addressed this problem by developing an ontology, called Medical WordNet, which aimed at bridging the language gap between the variations and vocabulary differences between experts and non-experts within the medical domain. This is to facilitate the development of NLP solutions and research within the domain. This approach could be explored within the context of VA.

10.5.5 Other Uses of the Verbal Autopsy Corpus

Adolphs et al.(2004) employed corpus based approaches and conversation analytics methods to explore the communications between health advisors and nurses and care seekers over telephone conversations within the United Kingdom's National Health System. This study helped in revealing operational issues regarding strategies employed by health advisors during interactions. The VA corpus allowed some of these approaches to be explored. This could serve as a means to identifying interactions between interviewers and respondents during VA data collection. For example cultural issues surrounding deaths could better be understood using corpus based methods.

10.6 Summary

In summary, this Chapter has given an overall summary of the motivation, aims and objectives of this research project as reported in this thesis. This research project has demonstrated the value of open narrative text, which has been ignored by automatic approaches to predicting cause of death from VAs till date. We also demonstrated the benefits of combining information from two sources: closed responses and open narratives text sections of VAs. This was a novel contribution to the research community and practitioners of VAs such as the World Health Organisation.

The Chapter also outlined a potential research direction that could be explored in order to improve on the current results. Finally, we also suggested other potential uses of the VA to answer interesting research questions that will again be beneficial to the users of VAs around the world.

List of References

Adolphs, S., Brown, B., Carter, R., Crawford, P. and Sahota, O. (2004). Applying Corpus Linguistics in a health care context. *Journal of Applied Linguistics*,1:9-28.

Agrawal, A. J. and Kakde, O. (2013). Semantic analysis of natural language queries using domain ontology for information access from database. *International Journal of Intelligent Systems and Applications*, 5(12):81-90.

Akbani, R., Kwek, S. and Japkowicz, N. (2004). Applying Support Vector Machines to imbalanced datasets. *Machine Learning Lecture Notes in Computer Science*. Springer 3201:39-50

Amor, N. B., Benferhat, S. and Elouedi, Z. (2004). Naive Bayes vs. Decision Trees in intrusion detection systems. *Association for Computing Machinery*,pp420-424.

Anthony, L. (2005) Antconc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. *Professional Communication Conference*, *Tokyo, Japan*. Institute of Electrical and Electronics Engineers, pp729-737.

Atla, A., Tada, R., Sheng, V. and Singireddy, N. (2011). Sensitivity of different Machine Learning algorithms to noise. *Journal of Computing Sciences in Colleges.* 26:96-103.

Atwell, E. S., Hughes, J., and Souter, D. C. (1994). Amalgam: Automatic Mapping Among Lexicogrammatical Annotation Models. *In: The Balancing Act: Combining Symbolic and Statistical Approaches to Language-Proceedings of the Association for Computational Linguistics Workshop*. Association for Computational Linguistics. pp21-20

Bailly-Bechet, M., Bradde, S., Braunstein, A., Flaxman, A., Foini, L., and Zecchina, R.(2009). Clustering with shallow trees. *Journal of Statistical Mechanics: Theory and Experiment,* 12:12010.

Banko, M. and Brill, E. (2001). Mitigating the paucity-of-data problem: exploring the effect of training corpus size on classifier performance for natural language processing. *Proceedings of the First International Conference on Human Language Technology Research,* San Diego: Association for Computational Linguistics.pp1-5

Bates, M. (1995). Models of natural language understanding. *Proceedings of The National Academy of Sciences*, 92(22):9977-9982.

Bellman, R. (1957). Dynamic programming. Princeton, NJ: Princeton University Press

Biber, D. and Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26:263-286.

Bilenko, M., Basu, S. and Mooney, R. J. (2004). Integrating constraints and metric learning in semi-supervised clustering. *Proceedings of The Twenty-First International Conference on Machine Learning*, *Banff, Alberta, Canada*: Association for Computing Machinery,pp11-19

Bontcheva, K., Cunningham, H., Roberts, I. and Tablan, V. (2010). Web-based collaborative corpus annotation: requirements and a framework implementation. *New Challenges for Natural Language Processing Frameworks, Valletta, Malta*. [Accessed 17th March 2015]. Available from: https://gate.ac.uk/sale/lrec2010/teamware/teamware-lrec10.pdf

Box, GP, Hunter, WG, and Hunter, JS (1978), Statistics for experimenters: an introduction to design, data analysis, and model building, *New York:* Wiley

Brants, T. (2000). TnT: A statistical Part-of-Speech tagger. *Proceedings of The Sixth Conference on Applied Natural Language Processing, Seattle*, *Washington*. Association for Computational Linguistics, pp224-231

Breiman, L. (2001). Random Forests. *Machine Learning*, 45:5-32.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: a case study in Part-of-Speech tagging. *Computational Linguistics*. 21:543-565.

Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76:503-514.

Byass, P., Fottrell, E., Huong, D. L., Berhane, Y., Corrah, T., Kahn, K. and Muhe, L. (2006). Refining a probabilistic model for interpreting Verbal Autopsy data. *Scandinavian Journal of Public Health*, 34:26-31.

Byass, P., Kahn, K., Fottrell, E., Collinson, M. A. and Tollman, S. M. (2010). Moving from data on deaths to public health policy in Agincourt, South Africa: approaches to analysing and understanding Verbal Autopsy findings. *Plos Medicine*, 7.

Campbell OM, Gipson R. (1993). National Maternal Mortality Survey, Egypt 1992–93. Report of findings and conclusions. Cairo: *Directorate of Maternal and Child Health Care*, Ministry of Health and Population

Chandramohan, D., Maude, G. H., Rodrigues, L. C., and Hayes, R. J. (1994). Verbal Autopsies for adult deaths: issues in their development and validation. *International Journal of Epidemiology*, 23(2):213-222.

Charikar, M. S. (2002). Similarity estimation techniques from rounding algorithms. In *proceedings of the 34th Annual Symposium on Theory of Computing*. Association for Computing Machinery, pp380-388

Chawla, N. V., Japkowicz, N. and Kotcz, A. (2004). Editorial: special issue on learning from imbalanced data sets. *Association for Computing Machinery SIGKDD Explorations.* Newsletter, 6:1-6.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*,16(1):22-29.

Clark, A. (2003). Pre-processing very noisy text. In *Proceedings of Workshop on Shallow Processing of Large Corpora*, *Lancaster.* [Accessed 17th March 2015]. Available from: http://bultreebank.s481.sureserver.com/SProLaC/paper02.pdf

Cohen, A. M. (2006). An effective general purpose approach for automated biomedical document classification. *Proceedings of AMIA Annual Symposium:* American Medical Informatics Association, pp161-165

Cohen, A. M. and Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6:57-71.

Cohen, K. B., Ogren, P. V., Fox, L. and Hunter, L. (2005). Corpus design for biomedical natural language processing. In P*roceedings of the ACL-ISMB workshop on linking biological literature, ontologies and databases: mining biological semantics*. Association for Computational Linguistics, pp38-45.

Danso, S., Atwell, E., Johnson, O., ten Asbroek, G., Edmond, K., Hurt, C., Hurt, L., Zandoh, C., Tawiah, C., Fenty, J., Amenga-E, S., Owusu-Agyei, S. and Kirkwood, B. R.(2013a). A semantically annotated corpus for automatic Verbal Autopsy Analysis. *ICAME Journal of the International Computer Archive of Modern English*, 37:37-70

Danso, S., Atwell, E. and Johnson, O. (2013b). A comparative study of Machine Learning methods for Verbal Autopsy text classification. *International Journal of Computer Science Issues*,10(2). [Accessed 17th March 2015]. Available from: http://arxiv.org/abs/1402.4380

Debole, F. and Sebastiani, F. (2003). Supervised term weighting for automated text categorization. In *Proceedings of the 2003 Association for Computing Machinery* Symposium on Applied Computing. Association for Computing Machinery, pp784-788.

Dodd, G. G. (1969). Elements of data management systems. *Association for Computing Machinery Computing Surveys*, 1:117-133.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*. 19:1-74.

Edmond, K. M., Quigley, M. A., Zandoh, C., Danso, S., Hurt, C., Agyei, S. O. and Kirkwood, B. R. (2008). Aetiology of stillbirths and neonatal deaths in rural Ghana: implications for health programming in developing countries. *Paediatric and Perinatal Epidemiology*, 22:430-437.

Elebro, K., Rööst, M., Moussa, K., Johnsdotter, S., and Essén, B. (2007). Misclassified maternal deaths among East African immigrants in Sweden. *Reproductive Health Matters*, 15(30):153-162.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A. S. T., Soderland, S., Weld, D. S. and Yates, A. (2005). Unsupervised named-entity extraction from the web: an experimental study. *Artificial Intelligence*, 165:91-134.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters,* 27:861-874.

Fielding, A. H. and Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation,* 24:38-49.

Firth, J.R. (1957). Papers in Linguistics 1934–51. Oxford University Press.

Flaxman, A. D., Vahdatpour, A., Green, S., James, S. L. and Murray, C. J. L. (2011). Random Forests For VA Analysis: multisite validation study using clinical diagnostic gold standards. *Population Health Metrics*, 9.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, 3:1289-1305.

Forman, G. (2004). A pitfall and solution in multi-class feature selection for text classification. *In proceedings of the twenty-first International Conference on Machine Learning*. Association for Computing Machinery, p38.

Fottrell, E., Byass, P., Ouedraogo, T., Tamini, C., Gbangou, A., Sombie, I., Hogberg, U., Witten, K., Bhattacharya, S., Desta, T., Deganus, S., Tornui, J., Fitzmaurice, A., Meda, N. and Graham, W. (2007). Revealing the burden of maternal mortality: a probabilistic model for determining pregnancy-related causes of death from Verbal Autopsies. *Population Health Metrics.* 5

Fottrell, E, and Byass P. (2010). Verbal Autopsy: methods in transition. *Epidemiologic Reviews,* 32(1):8-55.

Francis, W. N. and Kucera, H. (1979). Brown corpus manual. Letters to the Editor, *Department of Linguistics, Brown University, Providence,* 5,7.

Freeman, J. V., Christian, P., Khatry, S. K., Adhikari, R. K., Leclerq, S. C., Katz, J. and Darmstadt, G. L. (2005). Evaluation of neonatal Verbal Autopsy using physician review versus algorithm-based cause-of-death assignment in rural Nepal. *Paediatric and Perinatal Epidemiology*, 19:323-331.

Friedman, J. H. (1997). On Bias, Variance, 0/1—Loss, And The Curse-of-Dimensionality. *Data Mining and Knowledge Discovery*, 1:55-77.

Gabrilovich, E. and Markovitch, S. (2006) Overcoming the brittleness bottleneck using wikipedia: enhancing text categorization with encyclopaedic knowledge. *Association for the Advancement of Artificial Intelligence*, 6:1301-1306.

Gajalakshmi, V. and Peto, R. (2006). Commentary: Verbal Autopsy procedure for adult deaths. *International Journal of Epidemiology,* 35:748-750.

Gamon, M. (2004). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th International Conference on Computational Linguistics*. Association for Computational Linguistics, p841.

Garenne, M. and Fauveau, V. (2006). Potential and limits of Verbal Autopsies. *Bulletin of the World Health Organization*, 84:164-164.

Ge, G. and Wong, G. W. (2008). Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles. *BMC Bioinformatics*, 9:275.

Gentleman, R. and Carey, V. (2008). Unsupervised Machine Learning. *Bioconductor Case Studies*. Springer, pp137-157.

Ghahramani, Z. (2001). An introduction to hidden markov models and Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(01):9-42.

Glinz, M. (2007). On non-functional requirements. In *Proceedings of Requirements Engineering Conference*. Institute of Electrical and Electronics Engineers, pp 21-26.

Goweder, A. and De Roeck, A. (2001). Assessment of significant Arabic corpus. Arabic Language Processing: Status and Prospects. In *Proceedings of Arabic NLP Workshop at European Chapter of the Association for Computational Linguistics*. *Toulouse*, *France*. [Accessed 17th March 2015]. Available from: http://www.elsnet.org/arabic2001/goweder.pdf

Graesser, A., Mcnamara, D., Louwerse, M. and Cai, Z. (2004). Coh-metrix: analysis of text on cohesion and language. Behaviour Research Methods, 36:193-202.

Greenbaum, S. and Nelson, G. (1996). The international Corpus of English (ICE) project. *World Englishes*, 15:3-15.

Quigley, M. A. (2005). Commentary: Verbal autopsies—from small-scale studies to mortality surveillance systems. *International Journal of Epidemiology*, 34(5):1087-1088.

Han, J., Kamber, M. and Pei, J. (2006). *Data Mining: Concepts and Techniques*, San Francisco: Morgan Kaufmann Publishers

Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1-14

Harris, Z. S. (1951). *Methods in Structural Linguistics*. Chicago, IL, US: University of Chicago Press.

Hoey, M. (2007). Lexical Priming: a new theory of words and language. *Functions of Language* 14(2): 283-294

Hospedales, T. M., Gong, S., and Xiang, T. (2013). Finding rare classes: active learning with generative and discriminative models. *Knowledge and Data Engineering*, Institute of Electrical and Electronics Engineers Transactions, 25(2):374-386.

Hotho, A., Nurnberger, A. and Paaãb, G. (2005). A brief survey of Text Mining. *Journal for Computational Linguistics and Language Technology,* 20:19-62

Hripcsak, G. and Heitjan, D. F. (2002). Measuring agreement in Medical Informatics reliability studies. *Journal of Biomedical Informatics*, 35:pp99-110.

Huang, Z., Eidelman, V., and Harper, M. (2009). Improving a simple bigram HMM part-of-speech tagger by latent annotation and self-training. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Companion Volume: Short Papers, pp213-216.

Hurt, L., Ten Asbroek, A., Amenga-Etego, S., Zandoh, C., Danso, S., Edmond, K., Hurt, C., Tawiah, C., Hill, Z. and Fenty, J. (2013). Effect of Vitamin A supplementation on cause-specific mortality in women of reproductive age in Ghana: a secondary analysis from the Obaapavita trial. *Bulletin of the World Health Organization*, 91:19-27.

ICD-10 (1992). International statistical classification of diseases and related health problems. *Tenth Revision. Geneva*: World Health Organization

James, S. L., Flaxman, A. D. and Murray, C. J. (2011). Performance of the tariff method: validation of a simple additive algorithm for analysis of verbal autopsies. *Population Health Metrics*, 9:31.

Japkowicz, N. and Stephen, S. (2002). The class imbalance problem: a systematic study. *Intelligent Data Analysis*, 6:29-449.

Jensen, D. D. and Cohen, P. R. (2000). Multiple comparisons in induction algorithms. *Machine Learning*, 38(3):309-338.

Jiang, L., Wang, D., Cai, Z. and Yan, X. (2007). Survey of improving Naïve Bayes for classification. *Advanced Data Mining and Applications*, pp134-145.

Joachims, T. (1998). Text Categorization with support vector machines: learning with many relevant features. Springer Berlin Heidelberg

John, G. H. and Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. San Francisco*: Morgan Kaufmann Publishers.

Johansson, S., Atwell, E., Garside, R., & Leech, G. (1986). The tagged LOB corpus: users' manual. *Norwegian Computing Centre for the Humanities*, Bergen.

Joshi, R., Lopez, A. D., Macmahon, S., Reddy, S., Dandona, R., Dandona, L. and Neal, B. (2009). Verbal Autopsy coding: are multiple coders better than one? *Bulletin of The World Health Organization*, 87:1-57.

Jurafsky, D. and Martin, J. H. (2000). Speech and Language Processing, Pearson Education India.

Jurafsky, D., Martin, J. H., Kehler, A., Vander Linden, K. and Ward, N. (2000). Speech and Language Processing: An introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, MIT Press.

Kahn, K., Tollman, S. M., Garenne, M. and Gear, J. S. S. (2000). Validation and application of Verbal Autopsies in a rural area of South Africa. *Tropical Medicine and International Health*, 5:824-831.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C. and Murthy, K. R. K. (2001). Improvements To Platt's SMO Algorithm For SVM Classifier Design. *Neural Computation*, 13; pp637-649.

Kim, J. D., Ohta, T., Tateisi, Y. and Tsujii, J. (2003). Genia Corpus - A semantically annotated corpus for Bio-Text mining. *Bioinformatics*, 19:180-I182.

King, G. and Lu, Y. (2008). Verbal Autopsy methods with multiple causes of death. *Statistical Science*, 23:78-91.

King, G., Lu, Y. and Shibuya, K. (2010). Designing Verbal Autopsy Studies. *Population Health Metrics*, 8:p19.

Kirkwood, B. R., Hurt, L., Amenga-Etego, S., Tawiah, C., Zandoh, C., Danso, S., Hurt, C., Edmond, K., Hill, Z., Ten Asbroek, G., Fenty, J., Owusu-Agyei, S., Campbell, O. and Arthur, P.(2010). Effect of Vitamin A supplementation in women of reproductive age on maternal survival in Ghana (Obaapavita): a cluster-randomised, placebo-controlled trial. *The Lancet*, 375:1640-1649.

Kirkwood, B. R., Manu, A., Ten Asbroek, A. H., Soremekun, S., Weobong, B., Gyan, T., Danso, S., Amenga-Etego, S., Tawiah-Agyemang, C. and Owusu-Agyei, S. (2013). Effect of the Newhints home-visits intervention on neonatal mortality rate and care practices in Ghana: a cluster randomised controlled trial. The Lancet, 381(9884):2184-2192.

Knowles, G. and Zuraidah Mohd, D. (2003). Tagging a corpus of Malay texts, and coping with 'syntactic drift'. In *Proceedings of the Corpus Linguistics Conference, Lancaster, UK.* [Accessed 17th March 2015]. Available from: http://ucrel.lancs.ac.uk/publications/cl2003/papers/knowles.pdf

Kohavi, R.(1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In P*roceedings of 14th International Joint Conference on Artificial Intelligence*. [Accessed 17th March 2015]. Available from: http://frostiebek.free.fr/docs/Machine%20Learning/validation-1.pdf

Kong, E. B. and Dietterich, T. G.(1995). Error-correcting output coding corrects bias and variance. In *proceedings of International Conference on Machine Learning,. Tahoe City, California, USA*. [Accessed 17th March 2015]. Available from: http://www.diku.dk/OLD/undervisning/2002e/156/Lectures/Lecture11/error-correcting-output-coding.pdf

Kotsiantis, S., Zaharakis, I. and Pintelas, P. (2006). Machine Learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26:159-190.

Kotsiantis, S. B., Zaharakis, I. and Pintelas, P. (2007). Supervised machine learning: a review of classification techniques. *Informatica*, 31:249–268

Kudo, T. and Matsumoto, Y. A (2004). Boosting algorithm for classification of semi-structured text. *In Proceedings of Conference on Empirical Methods in Natural Language Processing*, *Dekai Wu,* Hong Kong. pp301-308.

Kuo, F. Y. and Sloan, I. H. (2005). Lifting the curse of dimensionality. *Notices of The American Meteorological Society,* 52:1320-1328.

Lakeland, C. and Knott, A.(2004). Implementing a lexicalised statistical parser. In *Proceedings of the Australasian Language Technology Workshop*. *Sydney* Australia, pp47-54.

Lamontagne, L., Marchand, M., Caropreso, M. and Matwin, S. (2006). Beyond the bag of words: a text representation for sentence selection. *Advances In Artificial Intelligence. Lecture Notes in Computer Science*. Springer Heidelberg. 4013:324-335.

Lan, M., Tan, C. L., Low, H. B. and Sung, S. Y. (2005). A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. In *Proceedingsof I14th International Conference on World Wide Web.* ACM pp1032-1033.

Lan, M., Tan, C. L., Su, J. and Lu, Y. (2009). Supervised And Traditional Term Weighting Methods For Automatic Text Categorization. *Pattern Analysis And Machine Intelligence*, Institute of Electrical and Electronics Engineers Transactions, 31:721-735.

Landis, J. R. and Koch, G. G. (1977). The Measurement Of Observer Agreement For Categorical Data. *Biometrics*, pp159-174.

Lang, D. (2007). Consultant report-natural language processing in the health care industry. *Cincinnati Children's Hospital Medical Center*, Winter.

Leaman, R., Wojtulewicz, L., Sullivan, R., Skariah, A., Yang, J. and Gonzalez, G. (2010). Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. *Association for Computational Linguistics*, pp117-125.

Leech, G. (1993). Corpus annotation schemes. *Literary and Linguistic Computing*. 8:275-281.

Leopold, E. and Kindermann, J. (2002). Text categorization with support vector machines. How to represent texts in input space? *Machine Learning*, 46:423-444.

Lewis, D. D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings of the 15th Annual International Conference on Research and Development in Information Retrieval. Copenhagen, Denmark*: Association for Computing Machinery, pp37-50

Lewis, D. D. and Jones, K. S. (1996). Natural language processing for information retrieval. *Communications of the Association for Computing Machinery*. 39:92-101.

Li, S., Xia, R., Zong, C. and Huang, C. R. (2009). A framework of feature selection methods for text categorization. *Association for Computational Linguistics*, pp692-700.

Li, Y., Bontcheva, K. and Cunningham, H. (2009). Adapting SVM for data sparseness and imbalance: a case study in information extraction. *Natural Language Engineering,* 15:241-271.

Liu, Y., Loh, H. T. and Sun, A. (2009). Imbalanced text classification: a term weighting approach. *Expert Systems with Applications*, 36:690-701.

Loper, E. and Bird, S. (2002). NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1. Philadelphia, Pennsylvania:* Association for Computational Linguistics, pp63-70.

Lawn, J. E., Wilczynska-Ketende, K., and Cousens, S. N. (2006). Estimating the causes of 4 million neonatal deaths in the year 2000. *International Journal of Epidemiology*, 35(3):706-718.

Manning, C. D., Schutze, H. and Mitcognet (1999). Foundations of Statistical Natural Language Processing, MIT Press.

Matsumoto, S., Takamura, H. and Okumura, M. (2005). Sentiment classification using word sub-sequences and dependency sub-trees. *Advances in Knowledge Discovery and Data Mining*, pp21-32.

McCallum, A. and Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. In *Proceedings* of the Association for the Advancement of Artificial Intelligence-*98 workshop on learning for text categorization*, Madison, Wisconsin, USA, pp41-48.

McEnery, T. and Wilson, A. (1996). Corpus Linguistics, *Edinburgh University Press.* Edinburgh.

Mckeown, K. R. and Radev, D. R. (2000). Collocations. *Handbook Of Natural Language Processing*. Marcel Dekker.

Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the Association for Computing Machinery, 38(11):39-41.

Mitchell, T. M. (1997). Machine Learning. Mcgraw Hill International.

Moschitti, A. and Basili, R. (2004). Complex linguistic features for text classification: a comprehensive study. *Advances in Information Retrieval, Lecture Notes in Computer Science*, 2997:181-196.

Murray, C., Lopez, A., Feean, D., Peter, S. and Yang, G. (2007). Validation of the symptom pattern method for analysing verbal autopsy data. *Plos Medicine*, 4:1739-1753.

Murray, C. J., Lopez, A. D., Black, R., Ahuja, R., Ali, S. M., Baqui, A., Dandona, L., Dantzer, E., Das, V. and Dhingra, U. (2011). Population health metrics research consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets. *Population Health Metrics*, 9: 27.

Newman, M. E. J. 2005. Power laws, pareto distributions and Zipf's Law. *Contemporary Physics*, 46: 323-351.

Nigam, K., McCallum, A. K., Thrun, S. and Mitchell, T. (2000). Text classification from labelled and unlabelled documents using EM. *Machine Learning*, 39: 103-134.

Nikfarjam, A. and Gonzalez, G. H. (2011). Pattern Mining for Extraction of Mentions of Adverse Drug Reactions From User Comments. *AMIA Annual Symposium Proceedings*. *American Medical Informatics Association*.2011: pp1019.

Oberlander, J. and Nowson, S. (2006) Whose Thumb Is It Anyway?: Classifying Author Personality From Weblog Text. *Proceedings Of The Coling/ACL on Main Conference Poster Sessions*. Association For Computational Linguistics, pp627-634.

Pakhomov, S., Shah, N., Hanson, P., Balasubramaniam, S. and Smith, S. A. (2008) Automatic quality of life prediction using electronic medical records. *American Medical Informatics Association*, pp545.

Pakhomov, S., Weston, S. A., Jacobsen, S. J., Chute, C. G., Meverden, R. and Roger, V. L. (2007). Electronic Medical Records for Clinical Research: Application To The Identification Of Heart Failure. T*he American Journal of Managed Care*, 13: pp281.

Paul, D. B., and Baker, J. M. (1992). The design for the Wall Street Journal-based CSR corpus. In Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics, pp357-362

Pang, B. and Lee, L. (2004). A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Barcelona, Spain*: Association for Computational Linguistics. [Accessed 17th March 2015]. Available from: http://aclweb.org/anthology/P/P04/P04-1035.pdf

Pang, B. and Lee, L. (2005). Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. Annual Meeting-*Association for Computational Linguistics,* p115.

Pearce, D. A (2002). Comparative evaluation of collocation extraction techniques. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation,* pp1530-1536.

Pearce, D. (2001). Using conceptual similarity for collocation extraction. In *Proceedings of the Fourth Annual* CLUK *Colloquium.* [Accessed 17th March 2015]. Available from: *http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=14F65DC73*

C4B4FB2403A43BA3A436637?doi=10.1.1.12.2977&rep=rep1&type=pdf

Pereira, F., Mitchell, T. and Botvinick, M. (2009). Machine Learning Classifiers and FMRI: a tutorial Overview. Neuroimage, 45:199-209.

Pestian, J. P., Brew, C., Matykiewicz, P., Hovermale, D., Johnson, N., Cohen, K. B. and Duch, W. (2007). A shared task involving multi-label classification of clinical free text. *Association for Computational Linguistics*, pp97-104.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers,* 10:61-74.

Quigley, M., Schellenberg, J. and Snow, R. (1996). Algorithms for verbal autopsies: a validation study in Kenyan children. *Bulletin of the World Health Organization*, 74:147-154.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora*. *Hong Kong*: Association for Computational Linguistics,pp1-6

Reeves, B. C., and Quigley, M. (1997). A review of data-derived methods for assigning causes of death from Verbal Autopsy data. *International Journal of Epidemiology*, 26(5):080-1089.

Rennie, J. D., Shih, L., Teevan, J. and Karger, D.(2003). Tackling the poor assumptions of Naive Bayes text classifiers. In *Proceedings of the Twentieth International Conference on Machine Learning, Washington DC*, [Accessed 17th March 2015]. Available from: http://www.aaai.org/Papers/ICML/2003/ICML03-081.pdf

Rifkin, R. and Klautau, A. (2004). In defense of one-vs-all classification. *The Journal of Machine Learning Research*, 5:101-141.

Riloff, E. (1995). Little words can make a big difference for text classification. In *Proceedings of the 18th International Association for Computing Machinery*, *Special Interest Group on Information Retrieval Conference on Research and Development In Information Retrieval*. Association for Computing Machinery, pp130-136.

Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I. and Setzer, A. (2009). Building a semantically annotated corpus of clinical texts. *Journal Of Biomedical Informatics*, 42:50-966.

Ruch, P., Baud, R. and Geissbühler, A. (2003) . Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. *Artificial Intelligence In Medicine*, 29:169-184.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management,* 24:513-523.

Samuel, A. L. (2000). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 44:206-226.

Scott, S. and Matwin, S. (1999). Feature engineering for text classification. In *Proceedings of the Sixteenth International Conference on Machine Learning*. [Accessed 17th March 2015]. Available from: http://comp.mq.edu.au/units/comp348/reading/scott99feature.pdf

Sebastiani, F. (2002). Machine learning in automated text categorization. Association for Computing Machinery Computing Surveys, 34:1-47.

Seretan, V., Nerima, L. and Wehrli, E.(2003). Extraction of multi-word collocations using syntactic bigram composition. In *Proceedings of the Fourth International Conference on Recent Advances in Natural Language Processing, 10-12, Borovets, Bulgaria*. pp424-431.

Smith, B. and Fellbaum, C. (2004). Medical WordNet: a new methodology for the construction and validation of information resources for consumer health. In Proceedings *of the 20th International Conference on Computational Linguistics. Geneva, Switzerland*. Association for Computational Linguistics. [Accessed 17th March 2015]. Available from: http://delivery.acm.org/10.1145/1230000/1220409/p371-smith.pdf

Smith, K., Megyesi, B., Velupillai, S. and Kvist, M. (2014). Professional language in Swedish clinical text: Linguistic characterization and comparative studies. *Nordic Journal of Linguistics*, 37(02):297-323.

Sokolova, M.,and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management, 45*:427-437

Soleman, N., Chandramohan, D. and Shibuya, K. (2005). Word Health Organisation technical consultation on Verbal Autopsy tools. *Geneva. Switzerland* [Accessed 17th March 2015]. Available from: http://www.who.int/healthinfo/statistics/mort_verbalautopsy.pdf

Soleman, N., Chandramohan, D. and Shibuya, K. (2006). Verbal Autopsy: current practices and challenges. *Bulletin of the World Health Organization*, 84:239-245

Spasic, I., Ananiadou, S., Mcnaught, J. and Kumar, A. (2005). Text mining and ontologies in biomedicine: making sense of raw text. *Briefings in Bioinformatics*, 6:239-251.

Sriurai, W., Meesad, P., and Haruechaiyasak, C. (2010). Improving web page classification by integrating neighbouring pages via a topic model. In *Proceedings of 10th International Conference on Innovative Internet Community Systems*. GI-Lecture. pp238-246

Sun, A. and Lim, E. P. (2001). Hierarchical text classification and evaluation. *Computer Society*, Institute of Electrical and Electronics Engineers, pp 521.

Tateisi, Y., Yakushiji, A., Ohta, T. and Tsujii, J. I. (2005). Syntax annotation for the GENIA corpus. In *Proceedings of Second International Joint Conference on Natural Language Processing*. Lecture Notes in Artificial Intelligence, Springer, 3651:222-227.

Telishevka, M., Chenet, L., and McKee, M. (2001). Towards an understanding of the high death rate among young people with diabetes in Ukraine. *Diabetic Medicine*, 18(1):3-9.

Tsuruoka, Y., Tateishi, Y., Kim, J. D., Ohta, T., Mcnaught, J., Ananiadou, S. and Tsujii, J. I. (2005). Developing a robust Part-of-Speech tagger for biomedical text. *Advances in Informatics*. Lecture Notes in Computer Science, Springer, 3746:382-392

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia*, *Pennsylvania*: Association for Computational Linguistics, pp417-424.

Valentini, G. and Dietterich, T. G. (2004). *Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods*. *The Journal of Machine Learning Research*, 5:725-775.

Vapnik, V. N. (1999). An overview of statistical learning theory. *Neural Networks*, Institute of Electrical and Electronics Engineers Transactions,10:988-999.

Wallis, S. and Nelson, G. (2001). Knowledge Discovery In Grammatically Analysed Corpora. *Data Mining and Knowledge Discovery*, 5:305-335.

Wang, P. and Domeniconi, C. (2008). Building semantic kernels for text classification using Wikipedia. In *Proceeding of the 14th Association for Computing Machinery International Conference on Knowledge Discovery and Data Mining.* Association for Computing Machinery, pp713-721

Weiss, G. M. (1995). Learning with rare cases and small disjuncts. In *Proceedings of International Conference on Machine Learning, Tahoe City, California, USA*. [Accessed 17th March 2015]. Available from: http://www.aaai.org/Papers/AAAI/2000/AAAI00-102.pdf

Wilks, Y. and Stevenson, M.(1998). Word sense disambiguation using optimised combinations of knowledge sources. In *Proceedings of the 17th International Conference on Computational Linguistics-*. Association for Computational Linguistics, 2:1398-1402

Witten, I. H. and Frank, E. (2005). Data mining: practical machine learning tools and techniques, Morgan Kaufmann.

I. Witten, G. Paynter, E. Frank, C. Gutwin, and C. Nevill-Manning (1999). KEA: practical automatic keyphrase extraction. In *Proceedings of the Fourth . Association for Computing Machinery Conference on Digital Libraries. . Berkeley, CA, USA*. Association for Computing Machinery, pp254–255

Winbo, I. G., Serenius, F. H., Dahlquist, G. G., and Källén, B. A. (1998). NICE, a new cause of death classification for stillbirths and neonatal deaths. International journal of epidemiology, 27(3):499-504

World Health Organisation (2004). World Report on Knowledge for Better Health: *Strengthening Health Systems,* Geneva: World Health Organization. [Accessed 17th March 2015]. Available from: http://www.who.int/rpc/meetings/en/world_report_on_knowledge_for_better_health2.pdf

World Health Organisation (2012). Verbal Autopsy Standards: 2012 World Health Organization Verbal Autopsy Instrument, *Geneva, World Health Organization.* [Accessed 17th March 2015]. Available from: *http://www.who.int/healthinfo/statistics/WHO_VA_2012_RC1_Instrument.pdf*

Worzel, W., Almal, A. and Maclean, C. (2007). Lifting the curse of dimensionality. *Notices of the American Mathematical Society.* 52:1320-1328

Wu, H. C., Luk, R. W. P., Wong, K. F. and Kwok, K. L. (2008). Interpreting TF-IDF term weights as making relevance decisions. *Association for Computing Machinery Transactions on Information Systems,* 26:1-37

Yahya, A. (1989) On the complexity of the initial stages of Arabic text processing. *Great Lakes Computer Science Conference*, *Kalamazzo, Michiga*. U.S.A

Youmans, G. (1991). A new tool for discourse analysis: the vocabulary-management profile. *Language*. 67(4):763-789

Zheng, Z., Wu, X. and Srihari, R. (2004). Feature selection for text categorization on imbalanced data. ACM SIGKDD Explorations Newsletter, 6. pp80-89.

List of Abbreviations

CSMF	Cause-Specific Mortality Fractions
CHERG	Neonatal Child Health Epidemiological Reference Group-
DT	Decision Tree
DWU	Discriminative Word Units
FRP	False Positive Rate
ICE	International Corpus of English
NB	Niave Bayes
NICE	Neonatal and Intrauterine Death Classification to Etiology
NTF	Normalised Term Frequency
PoS	Part-of-Speech
SMO	Sequential Minimal Optimisation
SVM	Support Vector Machine
CoD	Cause of Death
TFiD	Term Frequency Inverse Document Frequency
TF	Term Frequency
VA	Verbal Autopsy
WHO	World Health Organisation

Appendix A - Data Management Manual

This document shows the manual which provides detailed information regarding the system that was setup to manage data obtained from the surveillance system that was established as part of a large-scale epidemiological study – the ObaapaVita study. The ObaapaVita study required individuals to routinely collect data, including Verbal Autopsies from 200,000 participants over the course of approximately 10-years, which amounted to over one million records that were handled by the data management system.

Data Management System Manual for the

Data Management Team

Kintampo Health Reseach Centre

By

Samuel Danso

April 2005

Last updated: February 2008

Introduction

The Data Management System (DMS) is developed to handle complex data on around 200,000 women. The large number of different forms that are processed creates a huge potential for error if the data is not managed in an efficient manner. This document describes the DMS processes and functionalities developed to ensure quality data was obtained. Database schema of the DMS is as shown in Figure 1 below.

Database Schema of the DMS.

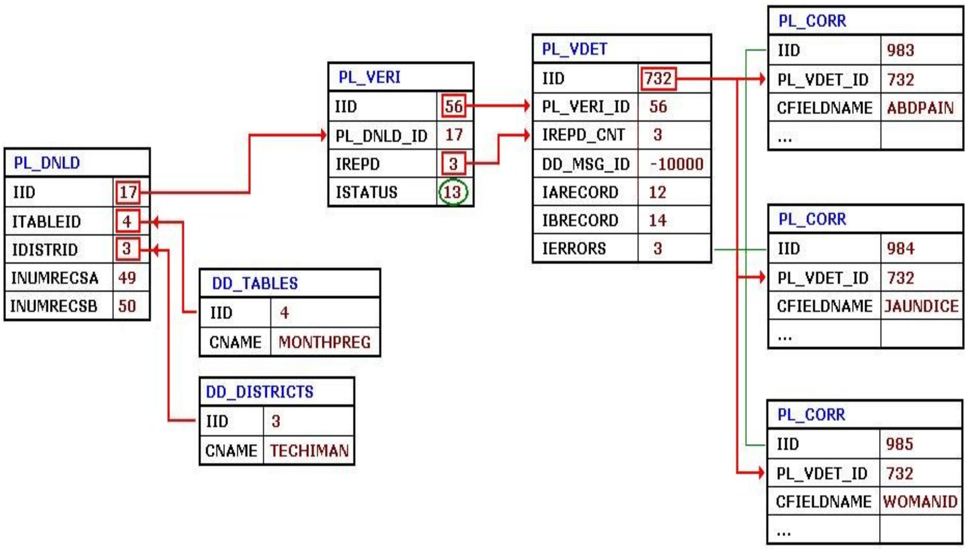


Figure 1. Database schema of the DMS

Figure 1 shows the database tables and relationships used in driving the system at various stages of the process.

Scheduling of fieldwork and distribution of forms

The DMS scheduling functionality is used to instruct each fieldworker which women they to visit each day and which forms need to be administered to each woman. The scheduling reports and forms are distributed from KHRC support office to the field site offices in the 7 districts in which the study was carried out, and the offices distributes the forms to fieldworkers via their field team coordinators and supervisors.

Data Entry

The DMS screens are designed to look the same as the fieldworkers' forms. The DMS runs on a network which is managed using a windows server. The protocol demands that each form is double entered by two different data entry clerks. Data entry occurs independently and verification checks are not performed as data was entered. This is because on-line verification was tried and was found to slow the process down too much to be feasible. The DMS temporally stores the data first for second streams of data entry to be stored on the server in sub-directory “A” and “B” respectively. These instructions are given to data entry clerks prior to entry. Data entry clerks work in teams to each clerk would always enter their data into the same stream. As a rule of thumb, data entry clerks should are to enter the data into the computer exactly as it appear on the form and must never change the information on the form. Data entry clerks do not specialize in entering particular forms. Each clerk enters all types of form. These forms are then stored in their respective tables within the DMS.

Data Verification

Data verification functionality of the DMD allows typographical errors to be corrected. This process detects errors made during data entry on computer but not errors made when the forms are completed. The data supervisors correct these errors and then repeat the verification check. After a verification check is performed the record is saved so that if necessary it can later be proved that the verification check was performed. Once the stream A and stream B entries are seen to be identical there should be no remaining errors created by mistakes in data entry. The data can then be passed on to the data managers for further cleaning.

Data Validation

Data validation functionality of the DMS allows range and logically inconsistent check to be performed. This is to ensure that the data is realistic. Range checks identify answers which are not possible. For example: one item on the Month questionnaire records how many capsules the woman in question has remaining from the last distribution. She could have 0, 1, 2, 3 or 4. These answers are acceptable, as is “8” to indicate the vial was not seen. The answer “9”, which is often used to denote not applicable is not acceptable as this question is always applicable. “5”, “6” or “7” are also impossible answers, as they fall outside the range of acceptable answers and thus are picked up with range checks. Logical inconsistencies are slightly different. Here is an example of a logical inconsistency:

In one question it is recorded that a woman answered that she had never given birth but in a subsequent question she answered that she has two children.

When some of these unacceptable entries or inconsistencies are identified, they complete query forms which are referred to the support office and subsequently to the field for checking and correction.

Database Updates and production of reports

At the end of each week the databases is updated. There are four databases which combine information from the questionnaire forms. These are:

- woman.dbf: lists all women enrolled in the study. This database is used for producing listings for the Month(standard), Month(pregnancy), 2-Week Postpartum, 6-Week Postpartum, Adult Verbal Autopsy, Adherence and Profile questionnaires.
- female10_14.dbf: lists women who are aged 10 to 14 years. This database can be used to update the women database as these women become old enough to enter the study.
- active_infants.dbf: lists infants who are currently being followed up. This database is use to produce listings for the Infant forms and will be used for the Infant Verbal Post-Mortem questionnaires.
- compounds.dbf: lists all compounds from the initial survey pre-recruitment. This database was used for performing geographical information system mapping to allocate compounds into clusters.

The DMS reports are generated based on these databases. This process is essential to ensure that every woman is visited at the correct time and that the correct questionnaires are administered. As the number of women in the study is extremely high the system for producing the reports must be robust. Various database management techniques are employed to efficiently manage this process within the DMS.

Additional progress reports and data analysis

An additional requirement of the DMS is to facilitate the analysis of the data collected and generate progress reports for key indicators. For flexibility the DMS system allow for the following options to be selected in terms of generating output: A Visual FoxPro Report, CSV (Comma Separated Values) file, Excel File. This is done each week and sometimes on ad-hoc basis as and when required. Reports include the following:

Information type	Details
1. Number of women visited	• Number (%) successfully seen • Number who died • Number of these deaths that occurred within the post-partum period • Number who became pregnant • Number of pregnancies ended • Number of births recorded
2. Number of babies born	• Number of deaths recorded
3. Number of infant visits made	• Number (%) successfully completed • Number of infant deaths
4. Summaries	• Registered in trial • Currently active • Died • Withdrew • Lost to follow-up (moved/prolonged absence) • Number of women moved within study area • Completed pregnancies • Registered births • Number of maternal deaths

Appendix B – The VA Questionnaire

This section provides a complete example of VA questionnaire used when collecting the data for this project. Personal details written on the questionnaire have been redacted for anonymity.

1. BACKGROUND and ID:

PLEASE VERIFY ALL THE INFORMATION ON THE PRINTED LABEL BELOW. START BY ASKING FOR THE MOTHER'S NEOVITA ID CARD. IF THE NEOVITA CARD IS MISSING, PLEASE ASK TO EXAMINE THE ANC CARD, THE CHILD HEALTH RECORD, or ANY OTHER SOURCE THAT MAY HAVE THE INFORMATION. IF ANY OF THE INFORMATION IS INCORRECT, PLEASE CORRECT IT.

1.1. Ncluster…………………………	1 4 2 5	NCLUSTER
1.2. Woman's ID…………	[redacted]	WOMANID
1.3. Woman's name……..	[redacted]	WOMNAME
1.4. Infant ID number………………	1 0 1 3 5 7	SUBJECTID
1.5. Date of delivery [080808 = NK]…………..	[redacted]	DATEDELIV
1.6. Date of death [080808 = NK]……………………	01 09 2010	DATEDIED

CONFIRM THE DATE OF DEATH LISTED ON THE LABEL. IF CONFIRM A DIFFERENT DATE, DOUBLE LINE THROUGH DATE ON LABEL AND RECORD CONFIRMED DATE.

1.6.1 Akwadaa no wu sutô berâ anaa pâ berâ?…………………… | 1. Rainy | 2. Dry | 8. NK | SEASON

1.6.2. IF DATE OF DEATH IS NK (08080808), then ask the woman: Na akwadaa no adi nfie sɛn bere wↄ wↄwuyɛ no?

1. Within 24 hours of birth	2. 1-13 days	3. 2-3 weeks		4. 4-7 weeks
5. 2-3 months	6. 4-5 months	7. 6-11 months	8. NK	9. NA, date known

AGED

1.7. PLACE THE INFANT IN ONE OF THE FOLLOWING GROUPS. CONFIRM THIS WITH THE RESPONDENT DURING THE INTERVIEW.

1. Stillbirth = Born dead / child did not cry or move or breathe after birth after 22w gestation	2. Early neonatal death = Live birth with age at death 0 to 6 days
3. Late neonatal death = Live birth with age at death 7-27 days	4. Postneonatal death = Live birth with age at death 28 days or more

CLASSIF

1.8. Did you verify the WOMAN'S ID on the label?	1. Yes, verified with Neovita card.	2. Yes, verified without Neovita card.	3. No, Neovita card lost and no other source to verify.

VERIFY

1.9. Date of interview: ……………………………	01 04 2013	DATEVISIT
1.10. Staff code: ……………………………………	J H	FW
1.11. Time interview began. 24 hour clock……………………	1 0 4 1	TIMEBEG
1.12. Time interview finished. 24 hour clock …………………	1 1 5 5	TIMEFIN

2. INFORMATION ABOUT THE RESPONDENT

ASK TO SPEAK TO THE MOTHER OR TO ANOTHER ADULT CARETAKER WHO WAS PRESENT DURING THE ILLNESS THAT LED TO THE INFANT'S DEATH IF THIS IS NOT POSSIBLE, ARRANGE A TIME TO VISIT THE HOUSEHOLD WHEN THE MOTHER OR CARETAKER WILL BE HOME.

2.1. IS A RESPONDENT AVAILABLE? ………………… | 1. Yes | 2. No | RESPOND

2.2. CONSENT GIVEN FOR VERBAL AUTOPSY?………… | 1. Yes | 2. No | 9. NA no respondent | VPMCONS

IF RESPONDENT DOES NOT GIVE CONSENT, OR THERE IS NO RESPONDENT, PLEASE TRY TO FIND ANOTHER RESPONDENT. IF NECESSARY PLEASE RETURN TO THE COMPOUND AT ANOTHER TIME TO FIND A RESPONDENT AND COMPLETE THE INTERVIEW. IF THERE WILL NEVER BE A RESPONDENT, PLEASE SUBMIT AS PROBLEM FORM.

2.3. RESPONDENT'S NAME: [redacted] RESPNAME

2.4. RESPONDENT'S AGE.......... | 7 | 1 | RESPAGE

2.5. WHAT IS THE RELATIONSHIP OF THE MAIN RESPONDENT TO THE DECEASED CHILD? RELATION

11. Mother	12. Father	13. Grandmother	14. Grandfather	15. Aunt	16. Uncle
17. TBA	18. Other male: ______		19. Other female : ______		

2.6. Na nipa a ôreyi nsâm yi ano no ne akwadaa no na âte aberâ a ôrebâwu no anaa?.......... | 1. Yes | 2. No | RESLIVE

2.7. TOTAL NUMBER PRESENT WHO PARTICIPATED AT INTERVIEW (EXCLUDING INTERVIEWER[S])?.......... | 0 | 2 | TOTINT

2.8. OF THOSE PARTICIPATING IN THE INTERVIEW, WERE THE FOLLOWING PERSONS PRESENT AT THE ILLNESS THAT LED TO STILLBIRTH, DEATH OR HOSPITALISATION?

2.8.1. The infant's mother..........	1. Yes	2 No	PRESDEATH1
2.8.2. The infant's father..........	1. Yes	2 No	PRESDEATH2
2.8.3. The infant's grandmother..........	1. Yes	2 No	PRESDEATH3
2.8.4. The infant's grandfather..........	1. Yes	2 No	PRESDEATH4
2.8.5. The infant's aunt..........	1. Yes	2 No	PRESDEATH5
2.8.6. The infant's uncle..........	1. Yes	2 No	PRESDEATH6
2.8.7. TBA..........	1. Yes	2 No	PRESDEATH7
2.8.8. Other, SPECIFY: ______	1. Yes	2 No	PRESDEATH8

2.9. IF THE RESPONDENT IS NOT THE MOTHER, GIVE THE REASON WHY: RESPMO

1. Mother is not resident compound member	2. Mother is not present	3. Mother is dead	4. Mother is not capable of answering	5. Mother refused
6. Other:		8. NK	9. NA, mother is informant	

2.10. Maame no apômuden te sɛn seisei? MHEALT

1. Healthy	2. Ill	3. Not alive	8. NK	9. Not applicable

2.11. IF THE MOTHER IS DEAD:

2.11.1. Nna dodoô sân wô awoô no akyi na ôwui?.......... | 9 | 9 | 9 | MODIED
[888 NK, 999 NA / DID NOT DIE, 000 if less than 24 hours or died during delivery]

2.11.2. Bosome dodoô sân wô awoô no akyi na ôwui?.......... | 9 | 9 | MODIEDMTH
[0 – 12; 88 = NK, 99 = NA / DID NOT DIE]

IF THE RESPONDENT IS THE MOTHER, SKIP QUESTIONS 2.12 AND 2.13 AND CONTINUE FROM SECTION 3

2.12. Sâ ânnyâ akwadaa no maame na âyi nsâm no ano a, nipa ôreyi nsâm no ano no adi mfie sân? [88=NK; 99 = NA; respondent is mother].......... | 7 | 1 | R_AGE

2.13. Âhee na wokôô sukuu duru a w'annkô bio? RESEDUL

11. None	12. Primary school	13. Middle, continuation school, JSS	14. Technical, commercial, SSS, Secondary school	
15. Post-middle college, secretarial	16. Post secondary, nursing, polytechnic	17. University	88. NK	99. NA, respondent is mother

3. INFORMATION ABOUT THE MOTHER

3.1 Na maame no adi mfie sɛn merɛ a akwadaa no wui no? (IN YEARS; 88=NK).......... | 2 | 0 | MAGE

3.2. Na maame no yâ ôman bân so nii anaa ôfiri ôman bân so?.......... | 1. Ghanaian | 2. Other (SPECIFY): | NATION

3.3. Na maame no yâ deân nii?

11. Akan: e.g. Bono, Ashanti, Fanti.etc.	12. Bimoda, Chokosi	13. Dagarti, Frafra, Kusasi	14. Fulani	GH_ETHNIC
15. Ga, Adangbe, Ewe	16. Gonja, Dagomba, Mamprusi	17. Konkomba,asare	18. Mo	
19. Sisala, Wala / 20. Zambraba	21. Banda/Pantra	22. Other, SPECIFY:		

3.4. Na maame no te mansin mua yeredi Neovita nhwehwe mu dwumadie no? .... | 1. Yes | 2. No | 8. NK | RESIDENCE

3.5. Mfie dodoô sân na ma | EDUWOMAN

3.6. Maame no kôô sukuu duru hee (anaa ôsua adeâ duru hee)?

11. None	12. Primary school	13. Middle,continuation school, JSS	14. Technical, commercial, SSS, Secondary school	EDUCLEVEL
15. Post-middle college, secretarial	16. Post secondary, nursing, polytechnic	17. University	88. NK	

3.7. Na maame no tumi kenkan nwoma anaa ôtumi twerâ? .... | 1. Yes | 2. No | 8. NK | READWRITE

3.8. Ansa na akwadaa no rebâwu no, na maame no yâ adwuma a âde sika berâ no?

11. Mainly employed	12. Mainly unemployed	14. Home maker	15. Student		OCCWOMAN
16. Pensioner	17. Does not work	18. Other	88. NK	99. NA	

If occupation of woman was 'other' then specify | MOCCOTH

3.9. Na maame no awaresâm te sân?..................

11. Never married / single	12. Married	13. Cohabiting		MARRIED
14. Separated	15. Divorced	16. Widowed	88. NK	

4. INFORMATION ABOUT THE CHILD

4.1. Akwadaa no yɛ baako anaa sɛ wôyɛ ntaafoo anaa ahenasa? .... | 1. Singleton | 2. Multiple | MULTIPLE

[IF TWO OR MORE CHILDREN ARE BORN, IT IS COUNTED AS A MULTIPLE BIRTH, EVEN IF ANY BABY IS BORN DEAD. IF MULTIPLE BIRTH THEN FILL A FORM FOR EACH BABY WHO DIES.]

4.2. Sɛ wôboro baako a, deɛ owuuiɛ no na ôdi kan anaa ɔtɔso mmienu anaa ɔtwa toɔ wɔ awoɔ no mu?

1. First	2. Second	3. Third or more	8. NK	9. Not applicable/singleton birth	BORDER

4.3. Na akwadaa no yɛ ôbaa anaa barima? .... | 1. Male | 2 Female | 8. NK | SEX

4.4. Na akwadaa no din de sɛn?
[Stillbirth = NA. FOR LIVE BIRTH USE DAY NAME IF NO OTHER NAME IS AVAILABLE] | CNAME

4.5. Mmere tenten sɛn na akwadaa no yare ansa na orewu?
[in days, 888 = NK, 000 = died during delivery; 999 = stillbirth]......................... | 0 | 0 | 0 | DURILL

* Died Sam day of deli

4.6. Akwadaa no wu mpofire mu anaa putupuru?.. | 1. Yes | 2 No | 8. NK | DIESUDD

JH

5. OPEN HISTORY QUESTIONS

5.1. Story of the illness

ALLOW THE RESPONDENT TO TELL YOU ABOUT THE PREGNANCY, DELIVERY AND THE BABY'S INJURY OR ILLNESS IN HER OWN WORDS. WRITE DOWN WHAT THE RESPONDENT TELLS YOU IN HER OWN WORDS. DO NOT PROMPT EXCEPT FOR ASKING WHETHER THERE WAS ANYTHING ELSE AFTER THE RESPONDENT FINISHES. KEEP PROMPTING UNTIL THE RESPONDENT SAYS THERE WAS NOTHING ELSE. WHILE RECORDING, UNDERLINE ANY UNFAMILIAR TERMS.

ALSO REMEMBER TO PROMPT ABOUT CARESEEKING DURING PREGNANCY, LABOUR, DELIVERY, AFTER THE BIRTH OF THE CHILD AND DURING THE FATAL ILLNESS. ASK WHAT THE MOTHER DID AND WHO SHE SOUGHT CARE FROM DURING ALL OF THESE TIMES.

FIRST ASK "Wobetumi aka biribi afa berɛ a na wonyem akwadaa no ho akyerɛ me?"

[redacted] trice never attended Antenatal care clinic (ANC) during her pregnancy.
She dose not felt sick during the pregnancy.
~~She also attended Antenatal care clinic (ANC)~~
~~for.~~ She ([redacted] delivere prematurely at seven (7) months.

THEN ASK "Wobetumi aka biribi afa berɛ a awoɔ kaa wo kɔpem sɛ wowooeɛ akyerɛ me?"

[redacted] complains of Lower Abdominal Pains (LAP) and stomach pains on the Wednesday when the the pregnancy was seven (7) months old.
She was taken to the [redacted] Hospital [redacted] on the Wednesday.
She had a normal delivery to a premature Female Baby (Seven (7) months), Alive but weak.

THEN ASK "Wobetumi aka biribi akyerɛ me afa sedeɛ na akwadaa no teɛ berɛ a wowoo no no?"

A Female premature Baby delivered looking smallish and also had difficulty in breathing. Also looking weak and could not cry at birth nor breastfeed.

THEN ASK "Wobetumi aka biribi akyerɛ me afa deɛ esii akwadaa no ho wɔ awoɔ no akyi?" [IN THIS QUESTION WE WANT TO KNOW WHAT HAPPENED TO THE BABY IMMEDIATELY AFTER DELIVERY. THAT IS IF THE BABY NEEDED ANY TREATMENT OR SPECIAL CARE AS SOON AS HE/SHE WAS BORN.']

The sick Newborn Baby was put into an incubator

THEN FOR LIVE BIRTHS ONLY ASK "Wobetumi aka biribi akyerɛ me afa yareɛ anaa nkwanhyia a ekuu akwadaa no ho?"
[IF STILLBIRTH PUT A DOUBLE LINE THROUGH THIS SECTION]

Fetal illness started from birth with difficulty in breathing.
She ([redacted]) expired at the [redacted] Hospital while still in the incubator.
She died within 24 hours of birth at 12:00AM. on ~~the~~ Thursday.

5.2 Cause of death
Wodwene sɛ ɛdeɛn na ɛde owuo no baeɛ?

5.2.1 Cause of death 1	88	COD1
5.2.2 Cause of death 2	99	COD2

6. DETAILS OF PREGNANCY, LABOUR AND DELIVERY

COMPLETE THIS SECTION FOR ALL INFANTS

PLEASE TELL THE RESPONDENT THAT YOU WOULD LIKE TO ASK SOME QUESTIONS CONCERNING THE MOTHER AND SYMPTOMS THAT THE DECEASED HAD/SHOWED AT BIRTH AND SHORTLY AFTER. PLEASE SAY THAT SOME OF THESE QUESTIONS MAY NOT APPEAR TO BE DIRECTLY RELATED TO THE BABY'S DEATH BUT ASK THE RESPONDENT TO PLEASE BEAR WITH YOU AND ANSWER ALL THE QUESTIONS. THE WILL HELP US TO GET A CLEAR PICTURE OF ALL POSSIBLE SYMPTOMS THAT THE DECEASED HAD.

6.1. Pregnancy

6.1.1. Maame no wo nkwadaa dodoô sân, sâ wokan wôn a ôwo twene ka ho a, ansa na ôrewo akwadaa yi?

1. None	2. One	3. Two	4. Three	5. Four	6. Five or more	8. NK	PARITY

6.1.2 Na maame no awô nteteâ paneâ no nyinaa bi ansa na ôreduru ne mpanin fie so a na n'yinsân wei ka ho?..	1. Yes	2. No	8. NK	VACCINATEFULL

6.1.3. Nteteâ oaneâ no dodoô sen na w'awô?........................	1. One	2. Two	3. Three	4. Four	5. Five or more	8. NK	VACCINDOSE

6.1.4. Maame no wôô asesene paneâ no bi wô ayinsân wei mu?..............................	1. Yes	2. No	8. NK	TTOXOID

6.1.5. Asensene panɛɛ dodoɔ sɛn na wowɔɔ no saa anyinsɛn no mu? [00 = NONE, 88 = NK, ASK TO SEE ANY MEDICAL RECORDS, YELLOW CARD]............	0 0	TETTOXD
6.1.6. Efirii sɛ yɛwoo wo, wowɔɔ asensene panɛɛ sɛn ansa na worenyinsɛn wanyinsɛn no? [00 = NONE, 88 = NK, ASK TO SEE ANY MEDICAL RECORDS, YELLOW CARD]...........	8 8	TETTOXB
6.1.7. Mprɛ dodoɔ sɛn na ôde n'anyinsɛn no kɔɔ ayaresabea maa wɔhwɛɛ anyinsɛn no? [00 = NONE, 88=NK, ASK TO SEE ANTENATAL CARE RECORD, EXCLUDE ILLNESS]	0 0	ANC

6.1.8. Saa ayinsân wei mu no, wo nyaa mogya borosoô?		1. Yes	2. No	8. NK	HIBPPREG
6.1.8.1. Wei sii bâyâ bosome miensa a âtwatoô ansa na wore wo?...............	1. Yes	2. No	8. NK	9. NA	LATEPREGBP
6.1.9. Abere a na woyim no, mogya tuu wo bâberee firi w'ase?..................................		1. Yes	2. No	8. NK	VAGBLEED
6.1.9.1. Wei sii wô bosome nsia edi kan wô ayinsân no mu anaa?	1. Yes	2. No	8. NK	9. NA	EARLYVAGBLEED
6.1.9.2. Wei sii wô bosome mneinsa âtwatoô wô ayinsân no mu anaa?.......	1. Yes	2. No	8. NK	9. NA	BLEED3MTHS
6.1.10. Wô bere a na wo yim no mu no, nsuo bi guu wo firi w'ase pân a na panpan wô mu?		1. Yes	2. No	8. NK	VAGDISC
6.1.10.1. Wei sii wô bosome miensa âtwatoô wô ayinsân no mu anaa? (Last 3 months)..	1. Yes	2. No	8. NK	9. NA	DISCHARGE
6.1.11. Wô w'ayinsân mu no, âdu bere bia na w'aniso yâ wo wisiwisi?......................		1. Yes	2. No	8. NK	BLURVIS

Ârebɛka bosome mmiensa ama woawo no wonyaa ôhaw a edidi soô yi bi? (Last 3 months)

6.1.12. Ntesheewa (Asthma)?..	1. Yes	2. No	8. NK	ASTHMA
6.1.13. Âtwerâ (Epilepsy)?..	1. Yes	2. No	8. NK	EPILEPSY
6.1.14. Kwashiokor (Malnutrition)?..	1. Yes	2. No	8. NK	MALNUT
6.1.15. Ôyââ kese boro so(Obesity)?..	1. Yes	2. No	8. NK	OBESITY
6.1.16. Kokoram yareâ (Cancer)?..	1. Yes	2. No	8. NK	CANCER
6.1.17. Nsamanwa (Tuberculosis)?..	1. Yes	2. No	8. NK	TB
6.1.18. Ahotutuo yareâ (Sickle cell disease)?..	1. Yes	2. No	8. NK	SICKLE
6.119. Akoma yareâ (Heart Disease)?..	1. Yes	2. No	8. NK	HEART_DIS
6.1.20. Wo homee a ânnsi so nso ânnyâ ntesheewa(Chronic obstructive pulmonary disease)?..	1. Yes	2. No	8. NK	COPD
6.1.21. (Dementia) (need to chwck with the doctors)?..	1. Yes	2. No	8. NK	DEMENTIA
6.1.22. Depression?.(same as above)..	1. Yes	2. No	8. NK	DEPRESS

6.1.23. Ne nipadua no fa dwodwoe (Stroke)?........		1. Yes	2 No	8. NK	STROKE
6.1.24. Ntini ntini ɓaa wo nan hō a na âyâ wo ya (Arthritis)?........		1. Yes	2 No	8. NK	ARTHRITIS
6.1.25. Saa yareâ (Kidney disease)?........		1. Yes	2 No	8. NK	KIDNEY
6.1.26. Berâboô yareâ (Liver disease)?........		1. Yes	2 No	8. NK	LIVER
6.1.27. Mogya tuu wo wɔ w'ase?........		1. Yes	2 No	8. NK	PBLEED
6.1.28. .Nsuo guu wo kâse wô w'ase a na ɛha adwene âsan sɛ na pampan wɔ mu anaa ɛdɔɔso dodo?........		1. Yes	2 No	8. NK	PDISCHARGE
6.1.29. 'Doctor' twee wo mogya hwɛe maa no ka kyerɛɛ wo sɛ wo mogya so ate?........		1. Yes	2 No	8. NK	PANAEMIA
6.1.30. 'Doctor' anaa nurse ka kyerɛɛ wo sɛ w'anya 'malaria'?........		1. Yes	2. No	8. NK	PMALARIA
6.1.31. 'Doctor' ka kyerɛɛ wo sɛ w'anya 'jaundice' (ɛda wani so sɛ akɔkɔ sradeɛ)?........		1. Yes	2 No	8. NK	PJAUNDICE
6.1.32. Wo yam yɛɛ wo ya kɛse paa anaa denden paa a ɛtoaso kyɛreeɛ aberɛ a na ɛnyɛ awoɔ na aka no?........		1. Yes	2 No	8. NK	PLONGPAIN
6.1.33. 'Doctor' anaa 'nurse' hwɛɛ no mogyaa maa no ka akyerɛ wo sɛ woanya asikyire yareɛ (diabetes)?........		1. Yes	2 No	8. NK	PDIABETES
6.1.34. Doctor anaa nurse ka kyerɛɛ wo sɛ woanya kekaeɛ?........		1. Yes	2 No	8. NK	PSYPHILIS
6.1.35. Wanyinsɛn no mu na ekuro wɔ wase a na ennwu da?........		1. Yes	2 No	8. NK	PULCER
6.1.36. Wo nsa anaa wanim honhonooe anaa sɛ wonan honhonoo ntɛmtɛn?........		1. Yes	2 No	8. NK	PFACE
6.1.37. Wani so yɛɛ wo basaa a na wohwɛ adeɛ a wohunu no mienu mienu ɛne woanya atipaeɛ a emu yɛ denden?........		1. Yes	2 No	8. NK	PBLUREYE
6.1.38. Tipaeâ denden?........		1. Yes	2 No	8. NK	HEADACHE
6.1.39. 'Doctor' susuu wo mogya maa no ka kyerɛɛ wo sɛ woanya mogya-boro soɔ?........		1. Yes	2 No	8. NK	PHIGHBP
6.1.40. Âsoro kaa wo sɛ nea ɛka nkwadaa no?........		1. Yes	2 No	8. NK	PCONVULSE
6.1.41. Wonyaa anidane wɔ anyinsɛn no mu?........		1. Yes	2 No	8. NK	PANIDANE
6.1.42. Afam /Atare?........		1. Yes	2 No	8. NK	PAFAM
6.1.43. N'anim honhonoe?........		1. Yes	2 No	8. NK	PALLOR
6.1.44. N'ani ase yɛɛ fitaa ne ne homee no nsisi so? (ne mienu no nyinaa wô hô)........		1. Yes	2 No	8. NK	PUFFY
6.1.45. Biribi foforô bi?........		1. Yes	2 No	8. NK	POTHER
6.1.45.1. Sâ biribi foforô biara, kyerâ mu [NA=DOUBLE LINE]........	//				OTHERSP_ILL

6.2. Labour and delivery

6.2.1. Wo woo akwadaa no bosome baako ansa na ne merâ âso?........		1. Yes	2 No	8. NK	PREMATURE
6.2.2. Wanyinsɛn no woberɛ a nnso, anaa wowoo wo berɛ ano, anaa wo berɛ paa ho?........	1. Early	2. On time	3. Late	8. NK	PREMBAB
6.2.3. Wonyinsɛn akwadaa no bosome sɛn? [88 = NK]........			0	7	GESTATE
6.2.4. Ârebɛka nna kakraa bi ama woawo no na akwadaa no keka ne ho? [9 = NA]...		1. Yes	2. No	8. NK	BABMOVE
6.2.5. Sɛ akwadaa no gyae sɛ ɔkeka ne ho a, abere bɛn na saa adeɛ yi si yɛ?........	11. Before labour started	12. During labour	88. NK	99 NA	MOVWHEN
6.2.5.1. Dônhwere dodoô sân ansa na worebâwo no na akwadaa no kekaa ne ho deâ âtwatoô? [00 = baby last moved during labour or delivery; 88 = not know]........			9	9	LASTMOVEHRS
6.2.6. Sɛ akwadaa no nkeka ne ho a, dabɛn na ɛtwa toɔ a wohunuu sɛ akwaada no akeka ne ho? [DAYS, 00 = baby last moved during labour or delivery; 88 = not known]			9	9	MOVDAY

6.2.7. Nsuo guu wase ansa na ɛregye wo akyem anaa ɛguu wo akyem no mu?

1. Before labour /delivery started	2. During labour / delivery	8. NK		WATERBRK

6.2.28. Hwan na ɔgyee wo awoɔ?

1. Doctor	2. Midwife or nurse	3. Trained TBA	7. Untrained TBA	6. Medical assistant	WHODELIV
5. Self	9. Relative	4. Other, SPECIFY:		8. NK	

6.2.29. Awoɔ ka no wɔ mu no, nea ɔgye awoɔ no tie sɛ akwadaa no akoma ɛbɔ/nakoma ɛyɛ adwuma?		1. Yes	2. No	8. NK	LISTEN
6.2.30. Sɛ aane a, na ɛbi wɔ hɔ/na ɛyɛ adwuma?	1. Yes	2. No	8. NK	9. NA	PRESENT

6.2.31. Na akwadaa no nna yafunu no mu yie (lying across or bottom first or breech) ansa na worebâwo no?	1. Yes	2 No	8. NK	POSITIONDEL

6.2.32. Akwadaa no nnipadua no fa hen na ɛdii kan baaeɛ?

1. Head	2. Bottom	4. Feet	5. Hand/arm	BPOSN
6. More than one body part (e.g. bottom and foot)		3. C-Section	8. NK	

6.3. Status of the baby at the delivery

6.3.1. Berɛ bɛn na wo woo akwadaa no wɔ da no mu?

1. 5am-7am	2. 8am-12pm	3. 1pm-4pm	4. 5pm-8pm	5. 9pm-1am	6. 2am-4am	TIMEDAY

6.3.2. Deɛn na wɔde twaa akwadaa no funuma?

1. Clinic/hospital instrument: scissors, razorblade, knife,etc	2. New razorblade/knife(not from clinic/hospital)		CORDCUT
3. Old razor blade/knife (not from clinic/hospital)	4. Other:	8. NK	

6.3.3. Wo woo akwadaa no na nkwa wɔ ne mu anaa na wawu?	1. Alive	2. Dead	8. NK	BORNALIVE
6.3.4. Wowoo akwadaa no ɔhomee da?	1. Yes	2. No	8. NK	BREATHB
6.3.5. Wɔyɔɔ biribi de boa akwadaa no sɛ ɔbɛtumi ahome?	1. Yes	2. No	8. NK	BREATHAS
6.3.6. Wo woo akwadaa no wieâ no, ôbrââ yâ ansa na ôrehome?	1. Yes	2. No	8. NK	DIFFBRATB
6.3.7. Wowoo akwadaa no ɔkekaa ne ho?	1. Yes	2. No	8. NK	MOVEB
6.3.8. Wowoo akwadaa no ɔsuu da?	1. Yes	2. No	8. NK	CRYB

6.3.8.1. Wowoo akwadaa no ɛdii mmerɛ dodoɔ sɛn na ɔhomee deɛ edi kan?

1. Within 5 min	2. Within 5-30 min	3. More than 30 min	8. NK	9. NA	FIRSTB

6.3.8.2. Wowoo akwadaa no ɛdii mmerɛ dodoɔ sɛn na ɔsuu deɛ edi kan?

1. Within 5 min	2. Within 5-30 min	3. More than 30 min	~~8. NK~~	9. NA	FIRSTCRY

6.3.9. Wowoo akwadaa no na ne kesɛɛ te sɛn? [PROMPT]

1. Tiny	2. Smaller than average	3. Average	4. Larger than most babies	5. Very big baby	8. NK	SIZE

6.3.10. RECORD BIRTHWEIGHT. [IN KILOGRAMS; 888 = NO RECORD] [ASK FOR ANC RECORD/DISCHARGE SLIP/WEIGHING CARD/HEALTH RECORD]	8 . 8 8	BIRTHWT

6.3.11. Wowoo akwadaa no na ne ho awɔtere anaa atiti?	1. Yes	2. No	8. NK	INJURY

6.3.12. Sɛ aane a, ɛhe na na nsensanyɛ anaa ahyensodeɛ a ɛkyerɛ sɛ w'apira anaa nnompe abubu no wɔ?

1. Head	2. Shoulder	3. Hips	SITEINJ
4. Face	5. Others, SPECIFY:	9. NA	

6.3.13. Na ahyensodeɛ bi d'adi sɛ akwadaa no abubu berɛ a wɔwoo no no?	1. Yes	2. No	8. NK	BIRPARAL
6.3.14. Wowoo akwadaa no na ne ho anoanoa anaa atwintwam anaa na ne ho ayɛ sɛ yɛamiamia no? (macerated, soft)?	1. Yes	2. No	8. NK	MACER
6.3.15. Wowoo akwadaa no na ne ho baabi adi dɛm?	1. Yes	2. No	8. NK	ANOMALY

6.3.15.1. Wowoo akwadaa no na ne ti sua paa?	1. Yes	2. No	8. NK	ANENC
6.3.15. 2. Wowoo akwadaa no, na ne ti so paa?	1. Yes	2 No	8. NK	HYDRO
6.3.15.3. Wowoo akwadaa no na ɛpɔ bi wɔ natikɔ anaa nakyi kasɛɛ mu a na ɛha adwene?	1. Yes	2.No	8. NK	BIFIDA
6.3.15.4. Wowoo akwadaa no na tokuro wɔ n'ano?	1. Yes	2. No	8. NK	CLEFT
6.3.15.5. Na ɛdɛm bi wɔ ne nan anaasɛ nensa?	1. Yes	2. No	8. NK	DEFECT
6.3.15.6. Na ɛdɛm foforɔ bi wɔ ne ho.	1. Yes	2.No	8. NK	OTHMALF
6.3.15.6.1. Sâ aane, kyerâ mu				OTHMALFSPEC

6.3.15.7. Na akwadaa no ahosuo te sɛn abere a wɔwo no no?

1. Normal	2. Pale	3. Blue	6. Other, SPECIFY:	COLBIRTH

6.3.15.8. Akwadaa no yâ bâtâá a na ônntumi nnkeka ne ho wô awoô no akyi (within 24 hours)?	1. Yes	2 No	8. NK	9. NA	UNCONS

7. NEONATAL AND POSTNEONATAL DEATHS – DETAILS OF FATAL ILLNESS

FOR STILLBIRTHS, PUT A DOUBLE LINE THROUGH SECTIONS 7-10 AND GO TO SECTION 11.

7.1. General

7.1.1. Da a wo woo akwadaa no na ne ho yɛ?		1. Yes	2 No	8. NK	DAYWELL
7.1.2. Yareɛ a ɛkuu no no hyɛɛ aseɛ no na wadi nna sɛn? [ANSWER EITHER 6.1.2.1. OR 6.1.2.2].					
7.1.2.1. IN DAYS [888 = NK, 999=NA, 000 = illness started at birth]		0	0	0	AGEILLD
OR					
7.1.2.2. IN MONTHS [88 = NK, 99=NA, 00 = illness started at birth]			9	9	AGEILLM
7.1.3. Akwadaa no wuuiɛ no na wadi nna sɛn? [ANSWER EITHER 6.1.3.1. OR 6.1.3.2].					
7.1.3.1. IN DAYS. [888 = NK, 999=NA]		0	0	0	AGEDIEDD
OR					
7.1.3.2. IN MONTHS [88 = NK, 99=NA]			9	9	AGEDIEDM
7.1.4. Wo woo akwadaa no ɔtumi suuiɛ?		1. Yes	2. No	8. NK	CRYNORM
7.1.5. ɛduruu bɛrɛ bi a na akwadaa no ntumi nsu?	1. Yes	2 No	8. NK	9. NA	CRYSTOP
7.1.6. ɔgyaee su no edii nna sɛn na ɔwuuiɛ? (in days) [000 = less than 1 day; 888 = not known; 999 = NA]		9	9	9	CRYDAYS
7.1.7. Na akwadaa no mfie a wadi kyeresɛ ɔrenyini sɛdeɛ ɛsɛ ?		1. Yes	2. No	8. NK	GROW
7.1.8. Na akwadaa no taa yare kyɛn nkwadaa nyinaa wɔ abusua no mu anaa borɔno so hɔ?		1. Yes	2. No	8. NK	SICKNESS

7.2. Feeding

7.2.1. Akwadaa no tumi nom nofoɔ da anaa odidi firi toa mu ne kuruwa anaa atere so?		1. Yes	2. No	8. NK	SUCKLE
7.2.2. Awoɔ no akyi no, edii merɛ sɛn ansa na akwadaa no nom nofoɔ, anaa odidi firi toa mu anaa kuruwa mu/kuruwa ne atere so?					
IN HOURS [888=NOT KNOWN 999=NA or > 1 day]		9	9	9	FHOURS
IN DAYS [888=NOT KNOWN 999=NA or < 1 day]		9	9	9	FDAYS
7.2.3. Akwadaa no gyae nofoɔ nom anaa ɔgyae sɛ odidi afiri toa mu anaa ɔgyae kuruwa ne atere so adidie?	1. Yes	2. No	8. NK	9. NA	SSUCK
7.2.4. Saa berâ no mu no, na akwadaa no tumi bue n'ano?	1. Yes	2 No	8. NK	9 NA	OPENMOUTH
7.2.5. Awoɔ no akyi no, edii merɛ sɛn ansa na akwadaa no gyae nofoɔ nom anaa ɔgyae toa mu aduane die anaa ɔgyae adidie wɔ kuruwa mu/kuruwa ne atere so?					
IN HOURS [888=NOT KNOWN 999=NA or > 1 day]		9	9	9	SHOURS
IN DAYS [888=NOT KNOWN 999=NA or < 1 day]		9	9	9	SDAYS

7.2.6. Ansa na akwadaa no bâwu no, berâ bân na ôgyae nom anaa adidie? IN DAYS 00 < 24 HRS; 88=NOT KNOWN 99=NA..	9	9			TIMENOSUCKLE
7.2.8. Na nofoɔ no nkoaa na ɔnom (me kyerɛ sɛ na wɔma akwadaa no nofoɔ nkoa a bibiara nnka ho gyesɛ vitamins, nnuro anaa ORS)?	1. Yes	2. No	8. NK	9. NA	EBF
7.2.9. Na akwadaa no yini sɛ nea ɛsɛ wɔ yadeɛ a ɛku no no mu?	1. Yes	2. No	8. NK		GROWTH
7.2.10. Ne so teeɛ yɛ?	1. Yes	2. No	8. NK		WTLOSS
7.2.11. Nna dodoɔ sɛn na ne so teeɛ? [IN DAYS, 000 = less than 1 day; 888 = not known; 999 = NA]..	9	9	9		WTLOSSD
7.2.12. Na w'afɔn ayɛ kitekite/ne mu atu?..	1. Yes	2. No	8. NK		FTHIN

7.3. Breathing

7.3.1. Akwadaa no nyaa yareɛ a ɛkuu no no na ɔbɔ wa?	1. Yes	2. No	8. NK		COUGH
7.3.2. Na ɛwa no ano yɛ den?	1. Yes	2. No	8. NK	9. NA	COUGHS
7.3.2.1. Na akwadaa no bɔ wa wieɛ a, ɔfe?	1. Yes	2. No	8. NK	9. NA	COUGHV
7.3.3. Awoɔ no akyi no, edii nna sɛn ansa na akwadaa no hyeaseɛ abɔ wa? [000 = less than 1 day; 888 = not known; 999 = NA]..	9	9	9		COUGHD
7.3.3.1. ɛwa no dii nna sɛn? (in days) 00 = less than 1 day; 88 = not known; 99 = NA]..	9	9			COUGHLEN

7.3.4. Na aka nna sân ama akwadaa no awu na ôhyââ aseâ sâ ôbôôwa?

1. On the same day of death	2. 1 to 3 days before death	3. 4 to 7 days before death	4. More than 1 week but within 1 month of death	5. More than 1 month before death	8. NK	9. NA	COUGHSTART

7.3.6. Berɛ a na ɔyare yareɛ a ɛkuu no no na ɔntumi nhome yiye?	1. Yes	2. No	8. NK	DIFFBR
7.3.7. Berɛ a na ɔntumi nhome yiye no na wadi nna sɛn? (in days) [000 = Just born or less than 1 day; 888 = not known; 999 = NA]..	0	0	0	DIFFBRDAY
7.3.8. Berâ tenten bân, ansa na ôrebâwu no na na ônntumi nnhome yiye? [00 less than 24 hours; 88 NK; 99 NA]..	0	0		TIMEDYSPN
7.3.9. Nna dodoɔ sɛn na w'antumi anhome yiye? (in days) [00 = less than 1 day; 88 = not known; 99 = NA]..	0	0		DIFFBRNUM
7.3.10. Berɛ a na ɔyare yareɛ a ɛkuu no no, na akwadaa no homee a ânnsi so?............	1. Yes	2 No	8. NK	BREATHLESS
7.3.10.1. Nna dodoô sân na na ôhomee a na ânnsi so (in days)?..	0	0		BRLESSDURATN
7.3.11. Yareɛ a ɛkuu no no mu na ɔhome ntɛmtɛm?	1. Yes	2. No	8. NK	FASTBR
7.3.12. ɔhyɛɛ aseɛ homee ntɛmtɛm no na wadi nna sɛn? (IN DAYS) [000 = Just born or less than 1 day; 888 = not known; 999 = NA]..	9	9	9	FASTBRDAY
7.3.13. Nna dodoɔ sɛn na ɔhomee ntɛmtɛm? (in days) [00 = less than 1 day; 88 = not known; 99 = NA]..	9	9		FASTBRNUM
7.3.14. Yareɛ a ɛkuu no no mu na ɔhome a ne koko tɔ mu ma wo hunu ne mfe?	1. Yes	2. No	8. NK	INDRAW
7.3.14.1. Bere tenten sɛn na ɔhomee a na ne mfe tɔ mu? (IN DAYS) [000 = less than 1 day; 888 = not known; 999 = NA]..	9	9	9	COUGHIN
7.3.15. Berɛ a na ɔyare yareɛ a ɛkuu no no, na akwadaa no bôwa a âgyegye ne mu?..	1. Yes	2 No	8. NK	WHOOP
7.3.16. ɔnyaa yareɛ a ɛkuu no no mu na ɔpene? [DEMONSTRATE]......................	1. Yes	2. No	8. NK	GRUNT
7.3.17. Yareɛ a ɛkuu no no mu na ɔhome a ne hwene ano bue bue?	1. Yes	2. No	8. NK	FLARE
7.3.18. Yareɛ a ɛkuu no no mu ɛbaa berɛ bi sɛ ogyaee home kyɛree na ɔsan homeeɛ?	1. Yes	2. No	8. NK	APNOEA
7.3.19. ɔnyaa yareɛ a ɛkuu no no akwadaa no nyaa muyareɛ?	1. Yes	2 No	8. NK	PNEU

7.4. Neurological problems

Question					Code
7.4.1. Yareɛ a ɛkuu no no mu ɛtwerɛ anaa ɛsoro kaa no?		1. Yes	2 No	8. NK	IFIT
7.4.1.1. Áwoɔ no akyi no, edii merɛ sɛn ansa na ɛsoro no ka akwadaa no? (IN DAYS) [000 = less than 1 day; 888 = not known; 999 = NA]		9	9	9	FITSTART
7.4.1.2. Âtwerɛ anaa ɛsoro kaa no no, edii nna sɛn ansa na ɛrekɔ? (IN DAYS) [00 = less than 1 day; 88 = not known; 99 = NA]			9	9	FITDAY
7.4.1.3. Âsoro kaa no no, amono mu hô ara na ne kôn kôô n'akyi (unconsciousness)?	1. Yes	2 No	8. NK	9. NA	FITCONSC
7.4.2. Akwadaa no yɛɛ winseen a nemu bu kôô akyire?		1. Yes	2. No	8.NK	ARCH
7.4.3. Yareɛ a ɛkuu no no mu, na pampam pagyaae?		1. Yes	2 No	8. NK	BULGE
7.4.3.1. Awoɔ no akyi no, edi merɛ sɛn ansa na n'apampam ɛpegya? [000 = less than 1 day; 898 = not known; 999 = NA]		9	9	9	BULGES
7.4.3.2. Napampam pegyae no, ɛdi nna dodoɔ sɛn ansa na orewu? [00 = less than 1 day; 88 = not known; 99 = NA]			9	9	BULGED
7.4.4. Yareɛ a ɛkuu no no mu, ne se ka sii so anaa ɔsensenee?		1. Yes	2No	8. NK	TETANUS
7.4.5. Akwadaa no nyaa yareɛ a ɛkuu no no na ɔntumi nkeka ne ho?		1. Yes	2 No	8. NK	COMA
7.4.5.1. Awoɔ no akyi no, edi merɛ sɛn ansa na akwadaa (bɔ ne din) no anntumi akeka ne ho?[000 = less than 1 day; 888 = not known; 999 = NA]		9	9	9	COMAS
7.4.5.2. Merɛ tenten sɛn na w'anntumi akeka ne ho? [00 = less than 1 day; 88 = not known; 99 = NA]			9	9	COMAL

7.4.5.3. Neho a na ɔnnkeka no baa prɛko pɛ wɔ dakorɔ mu anaa ɛbaa nkakrankakra wɔ nna beberee mu?

1. Suddenly	2. Over a single day	3. Slowly over many days	8. NK	9.NA	COMAON

7.5. Skin problems

7.5.1. Wowoo akwadaa no edii mmerɛ sɛn na wo anaa obi pepaa akwadaa no ho?

1. Within 30 minutes of birth	2. 30 minutes or later	3. The baby was never dried	8. NK	DRY

7.5.2. Wowoo akwadaa no edii mmerɛ sɛn na wo anaa obi de ntoma kyekyeree akwadaa no ho?

1. Within 30 minutes of birth	2. 30 minutes or later	3. The baby was never wrapped	8. NK	WRAP

7.5.3. ɛfirii sɛ wo woo no kɔsii sɛ ɔwuuiɛ, deɛn na wode kaa ne funuma so? NO THREAD, STRING, CLAMP

1. Nothing, left it alone	2. Hospital / clinic medicine	3. Shea butter	4. Leaves or herbs	5. Palm oil	CORDMED
6. Ground nut oil	7. Other:			8. NK	

Question					Code
7.5.4. Yareɛ a ɛkuu no no mu, ne funuma ano yɛɛ kɔkɔɔ anaa na ɛpu nsuo?		1. Yes	2. No	8. NK	DRAINUMB
7.5.5. Berɛ a na ôyare yareɛ a ɛkuu no no, ne funuma ano yâ kôkôô?		1. Yes	2 No	8. NK	UBMILICRED
7.5.5.1. Sâ aane a, na funuma no a ayâ kôkôô no terâ kôkaa n'ayaase honam no?	1. Yes	2 No	8. NK	9. NA	UMBILICSKINRED
7.5.5.2. Berɛ a na ôyare yareɛ a ɛkuu no no, na nsuo firi ne funuma no mu ba?		1. Yes	2 No	8. NK	UMBILICALPUS
7.5.6. Mo de biribi toto funuma no so wɔ awoɔ no akyi?		1. Yes	2. No	8. NK	APPLUMB

Sɛ aane a, na ɛyɛ deɛn a?

Question					Code
7.5.6.1. Chlorhexidine?	1. Yes	2. No	8. NK	9 NA	CHLOR
7.5.6.2. Gentian violet paint/ blue paint?	1. Yes	2. No	8. NK	9. NA	BLUE
7.5.6.3. Other, specify: ______	1. Yes	2. No	8. NK	9 NA	OTHAPPL

7.5.7. ɔnyaa yareɛ a ɛkuu no no, ɔnyaa neɛ edidi soɔ yi bi:

Question				Code
mpɔmpɔ a ahyehyɛ nsuo?	1. Yes	2. No	8. NK	PUSTULE
mfumfumgya?	1. Yes	2. No	8. NK	BLISTER
baabi kesɛɛ a ahyɛ nsuo?	1. Yes	2. No	8. NK	EMPYEMA
baabi a ahono na ayɛ kɔkɔɔ?	1. Yes	2. No	8. NK	ERYTHEMA
berɛ a ɔnyaa yareɛ a ɛkuu no no, ne honam baabi yɛɛ hye anaasɛ na ɛrehwanehwane?	1. Yes	2. No	8. NK	SRED

Ne honam ani baabi dane yââ tuntum?		1. Yes	2 No	8. NK	BLACKSKIN
7.5.8. Merâ tenten sân na ne honam no hyehyââ nsuo anaa baabi faako a âhyââ nsuo? [01 to 28, 88 NK, 99 NA]			9	9	PUSTULEDUR/

7.6. Diarrhoea and abdominal symptoms

7.6.1. Berɛ a na ôyare yareɛ a ɛkuu no no, na ôwô haw bi wô n'ayaase?		1. Yes	2 No	8. NK	ABDPROB
7.6.2. Akwadaa no nyaa yareɛ a ɛkuu no no na ɔkɔ tiefi nsuo nsuo a ɛmu yɛ merɛ ntɛmtɛm?		1. Yes	2. No	8. NK	STOOLS
7.6.3. Akwadaa no nyaa yareɛ a ɛkuu no no ônyaa ayamtuo?		1. Yes	2 No	8. NK	DIARR
7.6.4. Bâyâ nna sân ansa na ôrebâwu no na ayamtuo no hyââ aseâ? (IN DAYS) [00 = less than 1 day; 88 = NK; 99 = NA]		9	9	9	DIARRSTART
7.6.5. Ayamtuo no dii nna sɛn ansa na ɛregyae anaa ârekô? (IN DAYS) [000 = less than 1 day; 888 = not known; 999 = NA]		9	9	9	DIARRDAY
7.6.6. Da a ne yam tuu kesɛ paa no ɔkɔɔ tiefi dodoɔ sɛn? (IN DAYS) [88 = Not known, 99 = NA]			9	9	NUMDIARR
7.6.7. Wodwen sɛ wei kyerɛ sɛ akwadaa no tiafi nsuo nsuo no dɔɔso kyɛn sɛdeɛ ɛsɛ? [9 = NA]	1. Yes	2. No	8. NK	9. NA	DIARRNORM
7.6.8. Yareɛ a ɛkuu no no mu na mogya wɔ ne yamtuo no mu? [9 = NA]	1. Yes	2 No	8. NK	9. NA	BLDIARR
7.6.9 Berɛ a na ne yam tu no ɔnomm ORS, nsuo a yɛde nkyene agu mu nsuo bi a yɛakyɛrɛ no ayaresabea?	1. Yes	2. No	8. NK	9. NA	ORT
7.6.10. Ne yafunu worɔɔe?		1. Yes	2 No	8. NK	SWABDO
7.6.11. Awoɔ no akyi no, ɛdii nna sɛn na akwadaa no ayaase hyɛaseɛ sɛ ɛhonohono? [000=less than 1 day; 888=not known; 999=NA]		9	9	9	SWABDOD
7.6.12. Merɛ tenten sɛn na akwadaa no ayaase honhoneɛ? [00 = less than 1 day; 88 = not known; 99 = NA]			9	9	SWABDOL
7.6.13. Honohono no baa prɛko pɛ wɔ nna kakra ntam anaa nkakra nkakra wɔ bosome mu?	1. Rapidly over days	2. Slowly over months	8.NK	9.NA	SWABDOR
7.6.14. Biribi bɛhyɛɛ n'ayaase hɔ/biribi bɛfua n'ayaase hɔ?		1. Yes	2 No	8. NK	MASS
7.6.15. Merɛ tenten sɛn na biribi bɛhyɛɛ n'ayaase hɔ/biribi bɛfua hɔ? [00 = less than 1 day; 88 = not known; 99 = NA]			9	9	MASSL
7.6.16. Merɛ bi baa wɔ da no mu anaa nea ɛboro saa a akwadaa no anntumi annkɔ tiefi?		1. Yes	2 No	8. NK	CONSTIP
7.6.17. ɔnyaa yareɛ ɛkuu no no, na ɔfe biribiara?		1. Yes	2. No	8. NK	IVOMIT
7.6.17.1. Merɛ tenten sɛn wɔ awoɔ no akyi na akwadaa no hyɛ aseɛ sɛ ɔfe? [000 = less than 1 day; 888 = not known; 999 = NA]		9	9	9	WHOVOMIT
7.6.17.2. Feɛ no mu yɛɛ den paa no, mperɛ dodoɔ sɛn na na akwadaa no fe dakorɔ? [000 = less than 1 day; 88 = not known; 99 = NA]			9	9	WHOVDAY
7.6.18. Na akwadaa no feâ ani te sâ coffee" anaa kôkôô te sâ mogya ?		1. Yes	2 No	8. NK	VOMBLOOD

7.7 Injury

7.7.1. Akwadaa (bɔ ne din)no wuo no firi pira anaa nkwanhyia anaa aduronom anaa aboaka anaa gyahyeâ anaa nsuo fa no mu?		1. Yes	2. No	8. NK	INJURYI

IF THE INFANT DID NOT SUSTAIN AN INJURY THAT LED TO HER DEATH, DRAW A DOUBLE LINE THROUGH QUESTIONS 7.7.1.1 TO 7.7.15 AND MOVE STRAIGHT TO SECTION 7.8 OTHER PROBLEMS.

7.7.1.1. Obi n'âhyeda pira no anaa ɔmaa no nyaa akwanhyia no?	1. Yes	2 No	8. NK	9. NA	INTENTINJ
7.7.2. Na âyâ Car akwanhyia anaa?	1. Yes	2 No	8. NK	9.NA	RTA

IF NO DRAW A DOUBLE LINE THROUGH Q7.7.2.1 TO Q7.7.2.6.8 AND CONTINUE WITH Q7.7.3.

If yes:

7.7.2.1. Opiraaeâ esan sâ na ôte car ketewa no mu bi anaa?	1. Yes	2 No	8. NK	9.NA	CAR
7.7.2.2 Opiraaeâ âsan sâ na ôte Car keseâ (bus or heavy transport) no mu bi anaa?	1. Yes	2 No	8. NK	9.NA	BUS
7.7.2.3. Na ôte moto so anaa?	1. Yes	2 No	8. NK	9.NA	MOTO
7.7.2.4. Na ôte dadepônkô (bicycle) so anaa?	1. Yes	2 No	8. NK	9.NA	BIKE

7.4. Neurological problems

7.4.1. Yareɛ a ɛkuu no no mu ɛtwerɛ anaa ɛsoro kaa no?	1. Yes	2 No	8. NK		IFIT
7.4.1.1. Awoɔ no akyi no, edii merɛ sɛn ansa na ɛsoro no ka akwadaa no? (IN DAYS) [000 = less than 1 day; 888 = not known; 999 = NA]	9	9	9		FITSTART
7.4.1.2. Âtwerɛ anaa ɛsoro kaa no no, edii nna sɛn ansa na ɛrekɔ? (IN DAYS) [00 = less than 1 day; 88 = not known; 99 = NA]		9	9		FITDAY
7.4.1.3. Âsoro kaa no no, amono mu hô ara na ne kôn kôô n'akyi (unconsciousness)?	1. Yes	2 No	8. NK	9. NA	FITCONSC
7.4.2. Akwadaa no yɛɛ winseen a nemu bu kôô akyire?	1. Yes	2. No	8.NK		ARCH
7.4.3. Yareɛ a ɛkuu no no mu, na pampam pagyaae?	1. Yes	2 No	8. NK		BULGE
7.4.3.1. Awoɔ no akyi no, edi merɛ sɛn ansa na n'apampam ɛpegya? [000 = less than 1 day; 898 = not known; 999 = NA]	9	9	9		BULGES
7.4.3.2. Napampam pegyae no, ɛdi nna dodoɔ sɛn ansa na orewu? [00 = less than 1 day; 88 = not known; 99 = NA]		9	9		BULGED
7.4.4. Yareɛ a ɛkuu no no mu, ne se ka sii so anaa ɔsensenee?	1. Yes	2 No	8. NK		TETANUS
7.4.5. Akwadaa no nyaa yareɛ a ɛkuu no no na ɔntumi nkeka ne ho?	1. Yes	2 No	8. NK		COMA
7.4.5.1. Awoɔ no akyi no, edi merɛ sɛn ansa na akwadaa (bɔ ne din) no anntumi akeka ne ho?[000 = less than 1 day; 888 = not known; 999 = NA]	9	9	9		COMAS
7.4.5.2. Merɛ tenten sɛn na w'anntumi akeka ne ho? [00 = less than 1 day; 88 = not known; 99 = NA]		9	9		COMAL

7.4.5.3. Neho a na ɔnnkeka no baa prɛko pɛ wɔ dakorɔ mu anaa ɛbaa nkakrankakra wɔ nna beberee mu?

1. Suddenly	2. Over a single day	3. Slowly over many days	8. NK	9. NA	COMAON

7.5. Skin problems

7.5.1. Wowoo akwadaa no edii mmerɛ sɛn na wo anaa obi pepaa akwadaa no ho?

1. Within 30 minutes of birth	2. 30 minutes or later	3. The baby was never dried	8. NK	DRY

7.5.2. Wowoo akwadaa no edii mmerɛ sɛn na wo anaa obi de ntoma kyekyeree akwadaa no ho?

1. Within 30 minutes of birth	2. 30 minutes or later	3. The baby was never wrapped	8. NK	WRAP

7.5.3. ɛfirii sɛ wo woo no kɔsii sɛ ɔwuuiɛ, deɛn na wode kaa ne funuma so? NO THREAD, STRING, CLAMP

1. Nothing, left it alone	2. Hospital / clinic medicine	3. Shea butter	4. Leaves or herbs	5. Palm oil	CORDMED
6. Ground nut oil	7. Other:			8. NK	

7.5.4. Yareɛ a ɛkuu no no mu, ne funuma ano yɛɛ kɔkɔɔ anaa na ɛpu nsuo?	1. Yes	2. No	8. NK		DRAINUMB
7.5.5. Berɛ a na ôyare yareɛ a ɛkuu no no, ne funuma ano yâ kôkôô?	1. Yes	2 No	8. NK		UBMILICRED
7.5.5.1. Sâ aane a, na funuma no a ayâ kôkôô no terâ kôkaa n'ayaase honam no?	1. Yes	2 No	8. NK	9. NA	UMBILICSKINRED
7.5.5.2. Berɛ a na ôyare yareɛ a ɛkuu no no, na nsuo firi ne funuma no mu ba?	1. Yes	2 No	8. NK		UMBILICALPUS
7.5.6. Mo de biribi toto funuma no so wɔ awoɔ no akyi?	1. Yes	2. No	8. NK		APPLUMB

Sɛ aane a, na ɛyɛ deɛn a?

7.5.6.1. Chlorhexidine?	1. Yes	2. No	8. NK	9. NA	CHLOR
7.5.6.2. Gentian violet paint/ blue paint?	1. Yes	2. No	8. NK	9. NA	BLUE
7.5.6.3. Other, specify: ____________	1. Yes	2. No	8. NK	9. NA	OTHAPPL

7.5.7. ɔnyaa yareɛ a ɛkuu no no, ɔnyaa neɛ edidi soɔ yi bi:

mpɔmpɔ a ahyehyɛ nsuo?	1. Yes	2. No	8. NK	PUSTULE
mfumfumgya?	1. Yes	2. No	8. NK	BLISTER
baabi kesɛɛ a ahyɛ nsuo?	1. Yes	2. No	8. NK	EMPYEMA
baabi a ahono na ayɛ kɔkɔɔ?	1. Yes	2. No	8. NK	ERYTHEMA
berɛ a ɔnyaa yareɛ a ɛkuu no no, ne honam baabi yɛɛ hye anaasɛ na ɛrehwanehwane?	1. Yes	2. No	8. NK	SRED

Ne honam ani baabi dane yââ tuntum?......		1. Yes	2 No	8. NK	BLACKSKIN
7.5.8. Merâ tenten sân na ne honam no hyehyââ nsuo anaa baabi faako a âhyââ nsuo? [01 to 28, 88 NK, 99 NA]......			9	9	PUSTULEDURAT

7.6. Diarrhoea and abdominal symptoms

7.6.1. Berɛ a na ôyare yareɛ a ɛkuu no no, na ôwô haw bi wô n'ayaase?......		1. Yes	2 No	8. NK	ABDPROB
7.6.2. Akwadaa no nyaa yareɛ a ɛkuu no no na ɔkɔ tiefi nsuo nsuo a ɛmu yɛ merɛ ntɛmtɛm?......		1. Yes	2. No	8. NK	STOOLS
7.6.3. Akwadaa no nyaa yareɛ a ɛkuu no no ônyaa ayamtuo?......		1. Yes	2 No	8. NK	DIARR
7.6.4. Bâyâ nna sân ansa na ôrebâwu no na ayamtuo no hyââ aseâ? (IN DAYS) [00 = less than 1 day; 88 = NK; 99 = NA]......		9	9	9	DIARRSTART
7.6.5. Ayamtuo no dii nna sɛn ansa na ɛregyae anaa ârekô? (IN DAYS) [000 = less than 1 day; 888 = not known; 999 = NA]......		9	9	9	DIARRDAY
7.6.6. Da a ne yam tuu kesɛ paa no ɔkɔɔ tiefi dodoɔ sɛn? (IN DAYS) [88 = Not known, 99 = NA]......			9	9	NUMDIARR
7.6.7. Wodwen sɛ wei kyerɛ sɛ akwadaa no tiafi nsuo nsuo no dɔɔso kyɛn sɛdeɛ ɛsɛ? [9 = NA]......	1. Yes	2. No	8. NK	9. NA	DIARRNORM
7.6.8. Yareɛ a ɛkuu no no mu na mogya wɔ ne yamtuo no mu? [9 = NA]......	1. Yes	2 No	8. NK	9. NA	BLDIARR
7.6.9 Berɛ a na ne yam tu no ɔnomm ORS, nsuo a yɛde nkyene agu mu nsuo bi a yɛakyɛrɛ no ayaresabea?	1. Yes	2. No	8. NK	9. NA	ORT
7.6.10. Ne yafunu worɔɔe?......		1. Yes	2 No	8. NK	SWABDO
7.6.11. Awoɔ no akyi no, ɛdii nna sɛn na akwadaa no ayaase hyɛaseɛ sɛ ɛhonohono? [000=less than 1 day; 888=not known; 999=NA]......		9	9	9	SWABDOD
7.6.12. Merɛ tenten sɛn na akwadaa no ayaase honhoneɛ? [00 = less than 1 day; 88 = not known; 99 = NA]......			9	9	SWABDOL
7.6.13. Honohono no baa prɛko pɛ wɔ nna kakra ntam anaa nkakra nkakra wɔ bosome mu?......	1. Rapidly over days	2. Slowly over months	8.NK	9.NA	SWABDOR
7.6.14. Biribi bɛhyɛɛ n'ayaase hɔ/biribi bɛfua n'ayaase hɔ?......		1. Yes	2 No	8. NK	MASS
7.6.15. Merɛ tenten sɛn na biribi bɛhyɛɛ n'ayaase hɔ/biribi bɛfua hɔ? [00 = less than 1 day; 88 = not known; 99 = NA]......			9	9	MASSL
7.6.16. Merɛ bi baa wɔ da no mu anaa nea ɛboro saa a akwadaa no anntumi annkɔ tiefi?......		1. Yes	2 No	8. NK	CONSTIP
7.6.17. ɔnyaa yareɛ ɛkuu no no, na ɔfe biribiara?......		1. Yes	2. No	8. NK	IVOMIT
7.6.17.1. Merɛ tenten sɛn wɔ awoɔ no akyi na akwadaa no hyɛ aseɛ sɛ ɔfe? [000 = less than 1 day; 888 = not known; 999 = NA]......		9	9	9	WHOVOMIT
7.6.17.2. Fɛɛ no mu yɛɛ den paa no, mperɛ dodoɔ sɛn na na akwadaa no fe dakorɔ? [000 = less than 1 day; 88 = not known; 99 = NA]......			9	9	WHOVDAY
7.6.18. Na akwadaa no feâ ani te sâ coffee" anaa kôkôô te sâ mogya ?......		1. Yes	2 No	8. NK	VOMBLOOD

7.7 Injury

7.7.1. Akwadaa (bɔ ne din)no wuo no firi pira anaa nkwanhyia anaa aduronom anaa aboaka anaa gyahyeâ anaa nsuo fa no mu?......		1. Yes	2. No	8. NK	INJURY1

IF THE INFANT DID NOT SUSTAIN AN INJURY THAT LED TO HER DEATH, DRAW A DOUBLE LINE THROUGH QUESTIONS 7.7.1.1 TO 7.7.15 AND MOVE STRAIGHT TO SECTION 7.8 OTHER PROBLEMS.

7.7.1.1. Obi n'âhyeda pira no anaa ɔmaa no nyaa akwanhyia no?......	1. Yes	2 No	8. NK	9. NA	INTENTINJ
7.7.2. Na âyâ Car akwanhyia anaa?......	1. Yes	2 No	8. NK	9.NA	RTA

IF NO DRAW A DOUBLE LINE THROUGH Q7.7.2.1 TO Q7.7.2.6.8 AND CONTINUE WITH Q7.7.3.

If yes:

7.7.2.1. Opiraaeâ esan sâ na ôte car ketewa no mu bi anaa?......	1. Yes	2 No	8. NK	9.NA	CAR
7.7.2.2 Opiraaeâ âsan sâ na ôte Car keseâ (bus or heavy transport) no mu bi anaa?......	1. Yes	2 No	8. NK	9.NA	BUS
7.7.2.3. Na ôte moto so anaa?......	1. Yes	2 No	8. NK	9.NA	MOTO
7.7.2.4. Na ôte dadepônkô (bicycle) so anaa?......	1. Yes	2 No	8. NK	9.NA	BIKE

7.7.2.5. Na ôhyâ obi a ônam kwan ho akyi anaa?...	1. Yes	2 No	8. NK	9.NA	PED
7.7.2.6. Wonim biribi fa car no a ânyaa akwanhyia no anaa nipa car bôô no no anaa?..........	1. Yes	2 No	8. NK	9.NA	OTHVEH
Sâ aane, Na akwanhyia no fan ea âdidisoô yi bi ho:					
7.7.2.6.1. Nipa na ônam kwan no ho anaa kwan no nkyân?...........	1. Yes	2 No	8. NK	9.NA	OTHPED
7.7.2.6.2. Adeâ a na âgina anaa âsi faako?..................................	1. Yes	2 No	8. NK	9.NA	OTHSTAT
7.7.2.6.3.Kaa (Car) ?..	1. Yes	2 No	8. NK	9.NA	OTHCAR
7.7.2.6.4. Kaa keseâ anaa bus?..	1. Yes	2 No	8. NK	9.NA	OTHBUS
7.7.2.6.5. Moto?..	1. Yes	2 No	8. NK	9.NA	OTHMOTO
7.7.2.6.6. Dadepônkô (Bicycle) ?...	1. Yes	2 No	8. NK	9.NA	OTHBICY
7.7.2.6.7. Biribi foforô bi?	1. Yes	2 No	8. NK	9.NA	OTHELSE
7.7.2.6.8. Sâ aane a, kyerâ mu.					OTHELSE2
7.7.3. Ôpiraeâ a na ânnyâ kaa akwanhyia?......................................	1. Yes	2 No	8. NK	9.NA	NONRTA
7.7.4. Ôhweease anaa ôfiri tôô fam?...	1. Yes	2 No	8. NK	9.NA	FALL
7.7.5. Nsuo na âfaa no?..	1. Yes	2 No	8. NK	9.NA	DROWN
7.7.6. Obi na âde aduro guu n'duane mu? (poisoned)	1. Yes	2 No	8. NK	9.NA	POISONED
7.7. 7. Ôhyehyee na ôwui anaa?...	1. Yes	2 No	8. NK	9.NA	BURN
7.7.8. Obi na âboro no anaa?...	1. Yes	2 No	8. NK	9.NA	ASSAULT
7.7.8.1. Obi na âde tuo ku no anaa?..	1. Yes	2 No	8. NK	9.NA	FIREARM
7.7.8.2. Obi na ôde biribi ano yâ nam te sâ sekan na âkum no anaa?..............	1. Yes	2 No	8. NK	9.NA	STAB
7.7.8.3. Ôwui âsan sâ obi tataa no anaa ôboro no anaa?..............................	1. Yes	2 No	8. NK	9.NA	OTHASSAULT
7.7.8.3.1. Sâ aane a, kerâ mu........................					OTHASSAULT2
7.7.9. Aboa bi na âkaa n o anaa ntumoa (insects) na âwee no na âmaa no wui anaa?..........	1. Yes	2 No	8. NK	9.NA	BITE
7.7.9.1. Kraman na âkaa no ma no wui anaa?..	1. Yes	2 No	8. NK	9.NA	DOG
7.7.9.2. Ôwô na âkaa no ma no wui anaa?..	1. Yes	2 No	8. NK	9.NA	SNAKE
7.7.9.3. Ntumoa (insects) na âwee no ma no wui anaa?....................................	1. Yes	2 No	8. NK	9.NA	INSECT
7.7.9.4. Owui âsan sâ aboa bi na âkaa no anaa bôô no?.......................................	1. Yes	2 No	8. NK	9.NA	OTHBITE
7.7.9.4.1. Sâ aane a, kyerâ mu.......................					OTHBITE2
7.7.10. Ôpiraeâ âsan sâ biribi na âpim no anaa?..	1. Yes	2 No	8. NK	9.NA	NATURE
7.7.11. Afidie keseâ bi na âpir no anaa?..	1. Yes	2 No	8. NK	9.NA	MACH
7.7.12. Aboabi anaa adeâ bi na âpim no anaa?...	1. Yes	2 No	8. NK	9.NA	ANIM
7.7.13. Ade foforô bi na âpiraa no anaa?...	1. Yes	2 No	8. NK	9.NA	OTHINJ
7.7.13.1. Sâ aane a, kyerâ mu.......................					OTHINJ2

7.7.14. Pira no anaa aduronom anaa aboaka anaa gyahyeɛ anaa nsuo fa no akyi ɔdii nna sɛn na ɔwuuiɛ?

1. Died within 24 hours	2. Died 1 day later or more	3. Died at the site of the accident	8. NK	9. NA	DURINJ

7.7. 15. Akwanhyiâ no akyi no, ônyaa ayarehwâ (medical care) ansa na ôrewu?.................	1. Yes	2. No	8. NK	MDCARE

7.8. Other problems

Question					Code
7.8.1. ɔnyaa ahobene wɔ yareɛ a ɛkuu no no mu?		1. Yes	2. No	8. NK	IFEVER
7.8.1.1. Na ahobene/ahoɔhyeɛ no mu ɛden?	1. Yes	2. No	8. NK	9. NA	FEVSEV
7.8.1.2. Na ahobene/ahoɔhyeɛ no toa mu anaa ɛba a na ɛkɔ?	1. Yes	2. No	8. NK	9. NA	FEVINT
7.8.1.3. Awɔ te guu ne so anaa ne ho wosoeɛ?	1. Yes	2. No	8. NK	9. NA	RIGOR
7.8.2. Ahobene no hyɛɛ aseɛ na wadi nna sɛn? (IN DAYS) [000 = Just born or less than 1 day; 888 = not known; 999 = NA]		9	9	9	FEVERDAY
7.8.3. Ahobene no dii nna sɛn? (IN DAYS) [00 = less than 1 day; 88 = not known; 99 = NA]			9	9	FEVERNUM
7.8.4. Bâyâ nna sân ansa na ôbâwu no no ahobene anaa ahoôhyeâ no hyââ aseâ?	1. On the same day of death	2. 1 – 3 days before death	4. 4 – 7 days before death		FEVSTART
	4. More than one week before death		8. NK	9. NA	
7.8.5. Yareɛ a ɛkuu no no mu na wosɔ ne mu a ne ho yɛ nyunu?		1. Yes	2. No	8. NK	COLD
7.8.6. Ne ho hyɛɛ aseɛ yɛɛ nyunu no na wadi nna sɛn? (IN DAYS) [000 = Just born or less than 1 day; 888 = not known; 999 = NA]		9	9	9	COLDDAY
7.8.7. Nna dodoɔ sɛn na awɔ dee akwadaa no? (IN DAYS) [00 = less than 1 day; 88 = not known; 99 = NA]			9	9	COLDNUM
7.8.8. ɔnyaa yareɛ a ɛkuu no no ɔyɛɛ bɛtɛɛ bɛrɛ a na wadi agorɔ awie?		1. Yes	2. No	8. NK	LETHARGY
7.8.9. Yareɛ a ɛkuu no no mu ɔnyaa jaundice (ne honam anaa nani yɛɛ kɔkɔɔ sɛ akokɔ sradeɛ?		1. Yes	2. No	8. NK	IJAUND
7.8.9.1. Akwadaa no ani yɛɛ sɛ akokɔ sradeɛ?		1. Yes	2. No	8. NK	JAUNDE
7.8.9.2. Akwadaa no nsam anaa ne nan mu yɛɛ sɛ akokɔ sradeɛ?		1. Yes	2. No	8. NK	JAUNDS
7.8.9.3. Awoɔ no akyi no, ɛdii nna sɛn ansa na "jaundice" no hyeaseɛ? [000 = less than 1 day; 888 = not known; 999 = NA]		9	9	9	JAUNDB
7.8.9.4. "Jaundice" no di nna dodoɔ sɛn? [00 = less than 1 day; 88 = not known; 99 = NA]			9	9	JAUNDL
7.8.10. Yareɛ a ɛkuu no no mu nani yɛɛ kɔkɔɔ a na efiri nsuo tesɛ mpɔmpɔ mu nsuo?		1. Yes	2. No	8. NK	CONJUNCT
7.8.11. Yareɛ a ɛkuu no no mu mogya tu firii ne nipa dua mu baabi?		1. Yes	2. No	8. NK	HDN
7.8.12. Berɛ a na ôyare yareɛ a ɛkuu no no, n'apampam tôô mu (sunken fontannelle)?		1. Yes	2. No	8. NK	SUNKFONT
7.8.13. Na akwadaa no (bɔ ne din) wɔ yadeɛ koankrɔ anaa na ɔyare ansa na ɔnya akwanhyia anaa pira no?	1. Yes	2. No	8. NK	9. NA	INJURY2
7.8.13.1. Sɛ aane a, na ɛyɛ deɛn yadeɛ? (Please refer to code list)			9	9	CHRILL

3. POSTNEONATAL DEATHS - ADDITIONAL QUESTIONS ABOUT THE FATAL ILLNESS

COMPLETE THIS SECTION FOR POSTNEONATAL DEATHS ONLY.
FOR NEONATAL DEATHS PUT A DOUBLE LINE THROUGH THIS SECTION AND GO TO SECTION 9.
FOR STILLBIRTHS PUT A DOUBLE LINE THROUGH SECTIONS 8-10 AND GO TO SECTION 11.

8.1. Nutrition

Question				Code
8.1.1. ɔnyaa yareɛ a ɛkuu no no ɔfɔn yɛɛ kete kete?	1. Yes	2. No	8. NK	THIN
8.1.2. Bosome a ɔwuuiɛ no ɔfɔn yɛɛ kete kete?	1. Yes	2. No	8. NK	MARASMUS
8.1.3. Yareɛ a ɛkuu no no mu ne nan anaa ne nan ase honhonoeɛ?	1. Yes	2. No	8. NK	SWELL
8.1.4. Ahonhono no dii nna sɛn? (IN DAYS) [00= less than 1 day; 88 = not known; 99 = NA]				SWELLDAY
8.1.5. Abere a yadeɛ ɛku akwadaa no bɔ no no, n'anim honhonoeɛ?	1. Yes	2. No	8. NK	SFACE
8.1.6. Abere a yadeɛ ɛku akwadaa no bɔɔ no no, n'apɔ so honhoneɛ?	1. Yes	2. No	8. NK	SJOINTS
8.1.7. Abere a yadeɛ ɛku akwadaa no bɔɔ no no, ne nan pɔ so honhoneɛ?	1. Yes	2. No	8. NK	SANK
8.1.8. Abere a yadeɛ ɛku akwadaa no bɔɔ no no, ne (bɔ ne din) honam nyinaa honhoneɛ?	1. Yes	2. No	8. NK	SBODY
8.1.9. Ahonohono no dii mere sɛn ansa na ɛrekɔ? (IN DAYS) [00 = less than 1 day; 88= not known; 99 = NA]				SLAST

8.1.10. Yareɛ a ɛkuu no no mu ne honam hwanehwanee?	1. Yes	2. No	8. NK		FLAKE
8.1.11. Yareɛ a ɛkuu no no mu ne tiri nwi sesa yɛɛ kɔkɔɔ sɛ akokɔ sradeɛ?	1. Yes	2. No	8. NK		COLOR
8.1.11.1. Merɛ tenten sɛn na akwadaa no tiri nwii yɛ kɔkɔɔ anaa ne tiri nwii yɛ sɛ akokɔ sradeɛ? [00 = less than 1 day; 88 = NK; 99 = NA]					COLORD
8.1.12. ɔnyaa kwashiorkor wɔ bosome a ɛtwaam ansa na ɔrewu no mu?	1. Yes	2. No	8. NK		KWASHIOR
8.1.13. Yareɛ a ɛkuu no no mu ne mogya weeɛ anaa ɔyɛɛ fitaa?	1. Yes	2. No	8. NK		BABANAEM
8.1.14. Yareɛ a ɛkuu no no mu ne nsa mu yɛɛ fitaa?	1. Yes	2. No	8. NK		BABPALE
8.1.15. Yareɛ a ɛkuu no no mu ne mmɔwerɛ yɛɛ fitaa?	1. Yes	2. No	8. NK		BABWNAIL
8.1.16. Merɛ tenten sɛn na akwadaa no yɛ fitaa anaa ne nsam yɛ fitaa anaa ne mowerɛ so yɛ fitaa anaa n'ani yɛ fitaa? [00=less than 1 day;88=not known;99=NA]					BABANL

8.2 Breathing

8.2.1. ɔnyaa yareɛ a ɛkuu no no na ɔhome a ɛyɛ dede?	1. Yes	2. No	8. NK		NOISE
8.2.2. ɔnyaa yareɛ a ɛkuu no no na ɔhome kɔ ne mu a ne mene ase su? [DEMONSTRATE]	1. Yes	2. No	8. NK		STRIDOR
8.2.3. ɔnyaa yareɛ a ɛkuu no no na, ɔhome ba abɔnten a ne mene ase su? [DEMONSTRATE]	1. Yes	2. No	8. NK		WHEEZE

8.3 Neurological problems

8.3.1 Akwadaa no nyaa yareɛ a ɛkuu no no ne kɔn sensenee?	1. Yes	2. No	8. NK		STIFFNECK
8.3.1.1. Merâ tenten sân na ne kôn senseneneâ ? [00 = less than 1 day; 01 to 28; 88 = not known; 99 = NA]					STIFFNECKDUR
8.3.1.2. Akwadaa no nyaa yareɛ a ɛkuu no no na ɔntumi nsɔ adeɛ mu?	1. Yes	2. No	8. NK		GRASP
8.3.2. ɔgyae adeɛ mu sɔ no ɛdii nna sɛn na ɔwuuiɛ?	1. Less than 12 hours	2. 12 hours or more	8. NK	9. NA	STOPGRASP
8.3.3. Akwadaa no nya tipaeɛ?	1. Yes	2. No	8. NK		HACHE
8.3.4. Na tipaeɛ no mu ɛden?	1. Yes	2. No	8. NK	9. NA	HACHES
8.3.5. Tipaeɛ no dii merɛ tenten sɛn? [00 = less than 1 day; 88= not known; 99 = NA, no headache]					HACHEL
8.3.6. Akwadaa no fam (ɛfiri ne sisi/kɔpim ne nan ase) dwodwoeɛ?	1. Yes	2. No	8. NK		POLIO
8.3.7. Ndwodwoeɛ no hyeaseɛ preko pɛ wɔ dakorɔ mu anaa nkakrankakra wɔ nna beberee mu?	1. Suddenly / 2. Over a single day / 3. Slowly over many days		8. NK	9. NA	POLIOS
8.3.8. Akwadaa no nyaa yareɛ a ɛkuu no no, na yɛfrɛ no a ɔntumi nnye so?	1. Yes	2. No	8. NK		VOICE
8.3.9. ɔgyae nnyesoɔ no ɛdii nna sɛn na ɔwuuiɛ?	1. Less than 12 hours	2. 12 hours or more	8. NK	9. NA	STOPVOICE
8.3.10. Akwadaa no nyaa yareɛ a ɛkuu no no, na obi anaa biribi rekɔ a ɔntumi mfa nani nni n'akyi?	1. Yes	2. No	8. NK		FOLLOW
8.3.11. Akwadaa no gyae sɛ ɔde nani di biribi akyi no edii nna sɛn na ɔwuuiɛ?	1. Less than 12 hours	2. 12 hours or more	8. NK	9. NA	STOPFOLLOW

8.4 Skin problems

8.4.1. Bosome a ɛtwaa mu ansa na ɔrewu no mu ɔnyaa honam mu nsaa?	1. Yes	2. No	8. NK		RASH
8.4.2. Na nsaa no wɔ ne ho nyinaa?	1. Yes	2. No	8. NK	9. NA	GENRASH
8.4.3. Na nsaa no wɔ nanim?	1. Yes	2. No	8. NK	9. NA	FACRASH
8.4.4. Na nsaa no gu akwadaa no (bɔ ne din) mu? (ɛfiri ne kɔn mu kɔsi ne sisi)?	1. Yes	2. No	8. NK	9. NA	TRUNKRASH
8.4.5. Na nsaa no gu akwadaa (bɔ ne din) nsa ne ne nan ho?	1. Yes	2. No	8. NK	9. NA	ARMSRASH

8.4.6. Nsaa no dii nna sɛn? (IN DAYS) [00 = less than 1 day; 88 = not known; 99 = NA]........					RASHDAY
8.4.7. Na nsaa nó te sɛn?					
1. Measlés rash · 2. Rash with clear fluid · 3. Rash with pus			8. NK	9. NA	RASHAPP
8.4.8. Nsaa bi guu no a na nsuo a ani da hɔ wɔ mu?	1. Yes	2. No	8. NK	9. NA	BLISRASH
8.4.9. Nsaa no hyɛɛ aseɛ no ne honam paepaee anaa ehwanehwanee?	1. Yes	2. No	8. NK	9. NA	CRKRASH
8.4.10. Na saa nsaa yi yɛ ntoburo / ntɛnkyɛm?	1. Yes	2. No	8. NK	9. NA	MEASLES
8.4.11. Akwadaa no ani yɛ kɔkɔɔ?	1. Yes	2. No	8. NK	9. NA	REDEYES
8.4.12. Yareɛ a ɛkuu no no mu ne mmɔtoa mu honhonooe?	1. Yes	2. No	8. NK		AXLYMPH
8.4.13. Yareɛ a ɛkuu no no mu n'ahaa mu honhonooe?	1. Yes	2. No	8. NK		INGLYMPH
8.4.14. Berɛ a na ɔyare yareɛ a ɛkuu no no ne kɔn hohonooe?	1. Yes	2. No	8. NK		NECKLYMPH
8.4.15. Abere a yadeɛ ɛku akwadaa no bɔɔ no no, akwadaa no honam baabi honhonoeɛ?.	1. Yes	2. No	8. NK		OTHLYMPH
8.4.16. Merɛ tenten sɛn na ne honam no honhoeɛ? [00 = less than 1 day; 88= not known; 99 = NA]........					LYMPHL
8.4.17. Yareɛ a ɛkuu no no mu nsaa fitaa bi sii nano mu anaa ne keterɛma so?	1. Yes	2. No	8. NK		CANDIDA
8.4.18. Merɛ tenten sɛn na n'anum kuro no dii? (IN DAYS) [000 = less than 1 day; 888= not known; 999 = NA]........					CANDL

8.5 Abdominal problems

8.5.1. Akwadaa no nya yaew bi wɔ n'ayaase?	1. Yes	2. No	8. NK		APAIN
8.5.2. N'ayaase yaew no dii nna dodoɔ sɛn? (IN DAYS) [00 = less than 1 day; 88 = not known; 99 = NA]........					APAINL
8.5.3. Na n'ayaase yaew no mu yɛden?	1. Yes	2. No	8. NK	9. NA	APAINS

8.6 Other problems

8.6.1 Nsakraeɛ ba dwonsɔ dodoɔ a na akwadaa no dwonsɔ dabiara mu?	1. Yes	2. No	8. NK		URINE
8.6.2 Dwonsɔ dodoɔ sɛn na na akwadaa no dwonsɔ?					
1. Too much · 2. Too little · 3. No urine at all			8. NK	9. NA	URINEP
8.6.3. Nsakrayâ baa dwonsô dodoô a na ôdwonsô no dakorô no mu?........	1. Yes	2. No	8. NK		URINECHANGE
8.6.4. Merɛ tenten sɛn na dwonsɔ no ani sesa dea ɛtwa toɔ? [00 = less than 1 day; 88= not known; 99 = NA]........					URINEL
8.6.5. Âbaa sâ na kwadaa no ndwonsô koraa?........	1. Yes	2 No	8. NK		NOURINE
8.6.6. Berɛ a ôyare yarɛâ âtwatoô no, na ôdwonsô a mogya wô mu?........	1. Yes	2 No	8. NK		BLOODURINE
8.6.7. Akwadaa no ani keka kɔɔ mu abere a yadeɛ etwa toɔ bɔ no no?	1. Yes	2. No	8. NK		SUNK
8.6.8. Merɛ tenten sɛn na n'ani a ɛkeka kɔɔmu no dii? 00 = less than 1 day; 88= not known; 99 = NA]........					SUNKL
8.6.9. Mogya tuu akwadaa no firi ne hwenemu, n'anum anaa ne to?	1. Yes	2. No	8. NK		BLEED
8.6.10. Yareɛ a ɛkuu no no mu ɔnyaa 'malaria'?	1. Yes	2. No	8. NK		IMALARIA

9. INFORMATION ABOUT CARESEEKING

COMPLETE THIS SECTION FOR NEONATAL AND POSTNEONATAL DEATHS ONLY.
FOR STILLBIRTHS PUT A DOUBLE LINE THROUGH THIS SECTION AND SECTION 10 AND GO TO SECTION 11.

9.1. ɔnyaa yareɛ wei no wo anaa obi de no kɔhwehwɛɛ mmoa (baabi)?	1. Yes	2. No	8. NK	CARESEE
9.2. Wo anaa obi de saa yareɛ yi kɔɔ ayaresabea kɔpɛɛ ayaresa?	1. Yes	2. No	8. NK	APPCARE
9.3. Wo anaa obi de saa yareɛ yi kɔɔ dunsini hɔ kɔpɛɛ ayaresa?	1. Yes	2. No	8. NK	OTHCARE
9.4. Wɔde saa yareɛ/pira yi kɔɔ honhom mu nipa bi tese ɔsɔfoɔ, malam, bosomfoɔ, etc bi hɔ?	1. Yes	2. No	8. NK	RELICARE
9.5. Wɔde saa yareɛ /pira yi kɔ ɔyaresafoɔ bi a ɔte kuromu hɔ?	1. Yes	2. No	8. NK	COMMCARE
9.6. Wɔde saa yareɛ /pira yi kɔɔ dɔkota anaa nurse a ɔnoara yɔ n'adwuma hɔ?	1. Yes	2. No	8. NK	PPHYCARE
9.7. Obi a w'atete no sâ ôngye awoô wô fie (trained TBA)?..	1. Yes	2. No	8. NK	CSK_TBA
9.8. Homeopath?..	1. Yes	2. No	8. NK	CARESK_HMPTH
9.9. Wɔde saa yareɛ/pira yi kohwehwɛɛ mmoa baabi a yɛ tɔn nnuro?	1. Yes	2. No	8. NK	PHARMCARE
9.10. Wɔde saa yareɛ/pira yi kɔhwehwɛɛ mmoa wɔ obi a ɔnenam tɔn nnuro hɔ, sotɔɔ anaa ɛdwamu?	1. Yes	2. No	8. NK	DRUGCARE
9.11. Wɔde saa yareɛ/pira yi kɔhwehwɛ mmoa wɔ wo busuani anaa wadamfo hɔ?	1. Yes	2. No	8. NK	RELFRCARE
9.12. Wohwehwɛɛ moa firi 'clinic', 'health post' anaa ayaresabea akeseɛ mu fa saa yareɛ yi ho?	1. Yes	~~2. No~~	8. NK	FACCARE
9.12.1. Mprɛ dodoɔ sɛn?	0	1		FACCARENO
9.12.2. PLEASE PROVIDE FACILITY CODE1 [99=NA]........................	1	1		CODECARE1
9.12.3 PLEASE PROVIDE FACILITY CODE2 [99=NA]........................	9	9		CODECARE2
9.12.4 PLEASE PROVIDE FACILITY CODE3 [99=NA]........................	9	9		CODECARE3
9.12.5. PLEASE PROVIDE FACILITY CODE4 [99=NA]........................	9	9		CODECARE4
9.12.6 PLEASE PROVIDE FACILITY CODE5 [99=NA]........................	9	9		CODECARE5
9.12.7 PLEASE PROVIDE FACILITY CODE6 [99=NA]........................	9	9		CODECARE6
9.13. Wɔde saa yareɛ/pira yi kɔhwehwɛɛ mmoa baabi foforɔ bi? If yes, specify:____________	1. Yes	2. No	8. NK	OTHERCARE
9.14. Yareɛ no hyɛ aseɛ no edii nna sɛn na wohwehwɛɛ mmoa? [88=NK; 99=NA]...........	0	0		DCARS
9.15. Wô nna âtwatoô no mu no, wode akwadaa no kôô ayaresabea keseâ anaa ketewa mu?.....................	1. Yes	~~2. No~~	8. NK	TRAVELHOSP ne 15/04/13

9.16. Kwan bân so na wo fa kôô ayaresabea hô?...............

1. Walking / Foot	2. Bicycle	3. Motorbike	4. Car	5. Taxi
6. Bus / Tro-Tro	7. Other, SPECIFY: Baby was born and died at the Hospital			9.. NA

TRANSP * JH

9.17. Âhen anaa hwan hɔ na wohwehwɛɛ mmoa ɛdi kane? [USE FACILITY CODE LIST; 88=NK; 99=NA]..................	1	1	CARESEEK1
9.18. ɛhen anaa hwan hɔ na wohwehwɛɛ mmoa a etɔ so mmienu? [USE FACILITY CODE LIST; 88=NK; 99=NA]..................	9	9	CARESEEK2
9.19. ɛhen anaa hwan hɔ na wɔhwehwɛɛ mmoa a etɔ so mmiensa? [USE FACILITY CODE LIST; 88=NK; 99=NA]..................	9	9	CARESEEK3

9.20. Yareɛ a ɛkuu akwadaa no mu yɛgyee no too ayaresabea maa no daa hɔ da koro anaa sɛ ɛboroo saa?............	1. Yes	2. No	8. NK	PLACEADM
9.21. ɛhen na yɛgyee no tooe? [USE FACILITY CODE LIST; 99 = not admitted]..................	1	1		IHOSP

9.22. ɛhen naAKWADAA no wuuiɛ?

1. Clinic/hospital	2. Private maternity home	3. At home	PLACEDIED
4. On route to clinic/hospital	5. Other:		

Question					Code
9.23. IF THE ANSWER TO 8.18. IS 1 OR 2, STATE WHERE. [USE FACILITY CODE LIST; NA=99]......	1	1			ADDPLDIED
9.24. If the baby was discharged from hospital, what was their condition on discharge?......	1. Well	2. Somewhat unwell	3. Very unwell	9. NA	DISCHILL
9.25. Apɔmuden adwumayɛni no kaa dea ɛkuu no no kyerɛ wo?......	1. Yes	2. No	8. NK	9. NA	HWCOD

9.26. Deɛn na apɔmuden adwumayɛni no ka yɛ? HWSAY

9.27. Abere a wôgyee no too ayaresabea keseâ anaa ketewa mu no, wo nyaa ôhaw bi?......	1. Yes	2. No	8. NK	9. NA	ADMPROB
9.28. Wo nyaa ôhaw bi wô kwan a wôfaa so hwââ wo ba no (bô akwadaa no din) wô ayaresabea keseâ anaa ketewa mu no?......	1. Yes	2. No	8. NK	9. NA	TREATPROB
9.29. Wo nyaa haw bi kwan a wo bâfaso anya nnuro anaa wobâhwehwâ ne mogya mu wô ayaresabea keseâ anaa ketewa no mu?......	1. Yes	2. No	8. NK	9. NA	MEDPROB
9.30. Âfiri baabi a na akwadaa no te ârekô ayaresabea keseâ anaa ketewa mu no, na âdi bâyâ dônhwere mienu [Name's household]?......	1. Yes	2. No	8. NK	9. NA	DISTFAC
9.31. Wô nna âtwatoô no mu no, na moadwene yâ mo ntanta sâ akwadaa no bâhia ayarehwâ wô "doctor" hô (medical care was needed)?......	1. Yes	2. No	8. NK	9. NA	DOUBTS
9.32. Mo yââ abibi duro bi anaa?......	1. Yes	2. No	8. NK		TRADITIONMED
9.33. Abere a na ôrewu no, mo frââ obi wô telephone anaa mobile so sâ ômâboa mo?......	1. Yes	2. No	8. NK		CALLHELP
9.34. Abere a na akwadaa (bô ne din) no, âka mo bôô wô ne ho no sii mo kwan sâ mobâtua âfie hô ka foforô bi?......	1. Yes	2. No	8. NK		PROHIBCOST

10: TREATMENT AND RECORDS

COMPLETE THIS SECTION FOR NEONATAL AND POSTNEONATAL DEATHS ONLY. FOR STILLBIRTHS PUT A DOUBLE LINE THROUGH THIS SECTION AND GO TO SECTION 11.

10.1. Medicines

10.1.1. Berɛ a ɔyare no ɔnom aduro bi?......	1. Yes	2. No	8. NK	DRUGS

IF NO THEN PUT A DOUBLE LINE THROUGH THIS SECTION AND GO TO SECTION 10.2

Yareɛ a ɛkuu no no mu yɛmaa no nnuro a edidi soɔ yi bi?

10.1.2. Antibiotics......	1. Yes	2. No	8. NK	IANTIB
10.1.3. Aspirin......	1. Yes	2. No	8. NK	IASPIRIN
10.1.4. Anti-malarial......	1. Yes	2. No	8. NK	IANTIMAL

10.1.5. Atiridie (malaria) aduro bɛn na ɔnomeeɛ?......				ANTIMT
	1. Chloroquine	2. Fansidar	3. Quinine	
	4. Artesonate	5. Amodiaquine	8. NK	
	6. Artesonate-Amodiaquine		9. NA	
	7. Other, specify:			

10.1.6. Other known oral medicine......	1. Yes	2. No	8. NK	ORALMK
10.1.7. Other unknown oral medicine......	1. Yes	2. No	8. NK	ORALMUK
10.1.8. Antibiotic injection......	1. Yes	2. No	8. NK	ABINJ
10.1.9. Paneɛ foforɔ bi......	1. Yes	2. No	8. NK	OTHINJ
10.1.10. ORS......	1. Yes	2. No	8. NK	ORS
10.1.11. Wode nsuo sii ne so?......	1. Yes	2. No	8. NK	IVDRIP

10.1.12 Wɔbɔɔ no mogya........	1. Yes	2. No	8. NK	BLOODTRANS
10.1.13 Treatment / food through tube passed through nose........	1. Yes	2. No	8. NK	NGT
10.1.14 Biribi foforɔ, SPECIFY: ____	1. Yes	2. No	8. NK	OTHMED

10.2. Surgery

10.2.1. Woyɛ no operation esan yadeɛ no nti?	1. Yes	2. No	8. NK	OPER
10.2.2. Woyɛ no "operation" no, ɛdii nna dodoɔ sɛn ansa na orewu? [00 = less than 1 day; 88 = not known; 99 = NA]........	9	9		OPERL

10.2.3 Ne nipadua heefa na wɔyɛɛ "operation" no?

1. Abdomen	2. Chest	3. Head	4. Other, specify:	8. NK	9. NA	OPERP

10.2.4. Wôyââ akwadaa no "operation" foforô bi ansa na ôrewu?........	1. Yes	2. No	8. NK	SURGB4DEATHV

10.3 Immunisations

10.3.1. Wo ba no awɔ nsaman wa paneɛ? (BCG) [DEMONSTRATE WHERE THE IMMUNISATION IS INJECTED INTO TOP OF ARM]........	1. Yes	2. No	8. NK	BCG
10.3.2. Woba no wɔɔ ntɛnkyem paneɛ abere a onya bosome nsia kɔpim bosome dumienu no? [DEMONSTRATE WHERE THE IMMUNISATION IS INJECTED INTO THE ARM]........	1. Yes	2. No	8. NK	MEASIMM
10.3.3. Wô w'adwene mu no, akwadaa no nyaa nteteâ paneâ biara a na merâ aso sâ ônya bi ansa na ôrewu?........	1. Yes	2. No	8. NK	VACC

10.4 Health records

10.4.1. Akwadaa no wɔ doctor nkrataa ne 'weighing card'?	1. Yes	2 No	8. NK	IRECORD
10.4.2. Metumi atwerɛ deɛ ɛwɔ ne card anaa ne krataa no so agu me deɛ no so?	1. Yes	2 No	9. NA	ITRANSC

10.4.3. What type of health records does the respondent have?

Child health record / weighing card........	1. Yes	2. No	8. NK	HTYPE1
Mother's ANC card........	1. Yes	2. No	8. NK	HYTPE2
Burial permit........	1. Yes	2. No	8. NK	HYTPE3
Hospital prescription........	1. Yes	2. No	8. NK	HYTPE4
Treatment card........	1. Yes	2. No	8. NK	HYTPE5
Postmortem result........	1. Yes	2. No	8. NK	HYTPE6
Hospital discharge card	1. Yes	2. No	8. NK	HYTPE7
Laboratory results	1. Yes	2. No	8. NK	HYTPE8
Other documents, SPECIFY: ____	1. Yes	2. No	8. NK	HYTPE9

IF THERE ARE NO HEALTH RECORDS PUT A DOUBLE LINE THROUGH THIS SECTION AND SECTION 10.5 AND GO TO SECTION 10.6

10.4.4. TRANSCRIBE ALL THE ENTRIES WITHIN THE 12 MONTHS BEFORE THE CHILD DIED IF RESPONDENT ALLOWS YOU TO SEE THE RECORDS. INCLUDE ALL DATES.

MAKE SURE YOU INCLUDE ALL IMMUNISATIONS. WRITE THE DATE OF THE LAST MEDICAL NOTE IN SECTION 10.4.5.

IENTRY

10.4.5. RECORD THE DATE OF THE LAST MEDICAL NOTE [090909 = no note]........ IMEDNOTE

10.5. Infant weight

10.5.1. RECORD THE TWO MOST RECENT WEIGHTS OF THE INFANT IN KILOGRAMS

No date = 080808, No weight = 88.88

DO NOT INCLUDE BIRTHWEIGHT. BIRTHWEIGHT SHOULD BE INCLUDED IN SECTION 5. THE EARLIER ONE SHOULD BE DATE 1 AND THE LATER ONE SHOULD BE DATE 2 E.G. DATE 1 = 10/01/03, DATE 2 = 10/02/03.

Date 1									DATE1
Weight 1			.						WEIGHT1
Date 2									DATE2
Weight 2			.						WEIGHT2

10.6. Death certificate

10.6.1. Wowɔ ne wuo no ho andansedie krataa?	1. Yes	2. No	8. NK	DEATHC

ASK TO SEE THE DEATH CERTIFICATE AND RECORD WHETHER YOU ARE ABLE TO SEE IT.

10.6.2. ABLE TO SEE DEATH CERTIFICATE?	1. Yes	2. No	9. NA	VIEWDC

DRAW A LINE THROUGH THIS SECTION IF NO CERTIFICATE

10.6.3. RECORD THE IMMEDIATE CAUSE OF DEATH FROM THE CERTIFICATE — IMMCOD

10.6.4. RECORD THE FIRST UNDERLYING CAUSE OF DEATH FROM THE CERTIFICATE — UNDCOD1

10.6.5. RECORD THE SECOND UNDERLYING CAUSE OF DEATH FROM THE CERTIFICATE — UNDCOD2

10.6.6. RECORD THE THIRD UNDERLYING CAUSE OF DEATH FROM THE CERTIFICATE — UNDCOD3

10.6.7. RECORD THE CONTRIBUTING CAUSE(S) OF DEATH FROM THE CERTIFICATE — CONTCOD

11. ALCOHOL AND TOBACCO USE

11.1. Akwadaa (bô ne din) no maame nom nsa pân?	1. Yes	2. No	8 NK	ALCOHOL
11.2. Akwadaa no maame nom cigarette anaa taa da (cigarette, cigar, pipe etc.)?	1. Yes	2. No	8 NK	TOBACCCO

12. HIV/AIDS AND TUBERCULOSIS

SAY "THE FINAL QUESTIONS ABOUT THE BABY'S DEATH ARE ABOUT HIV/AIDS AND TB"

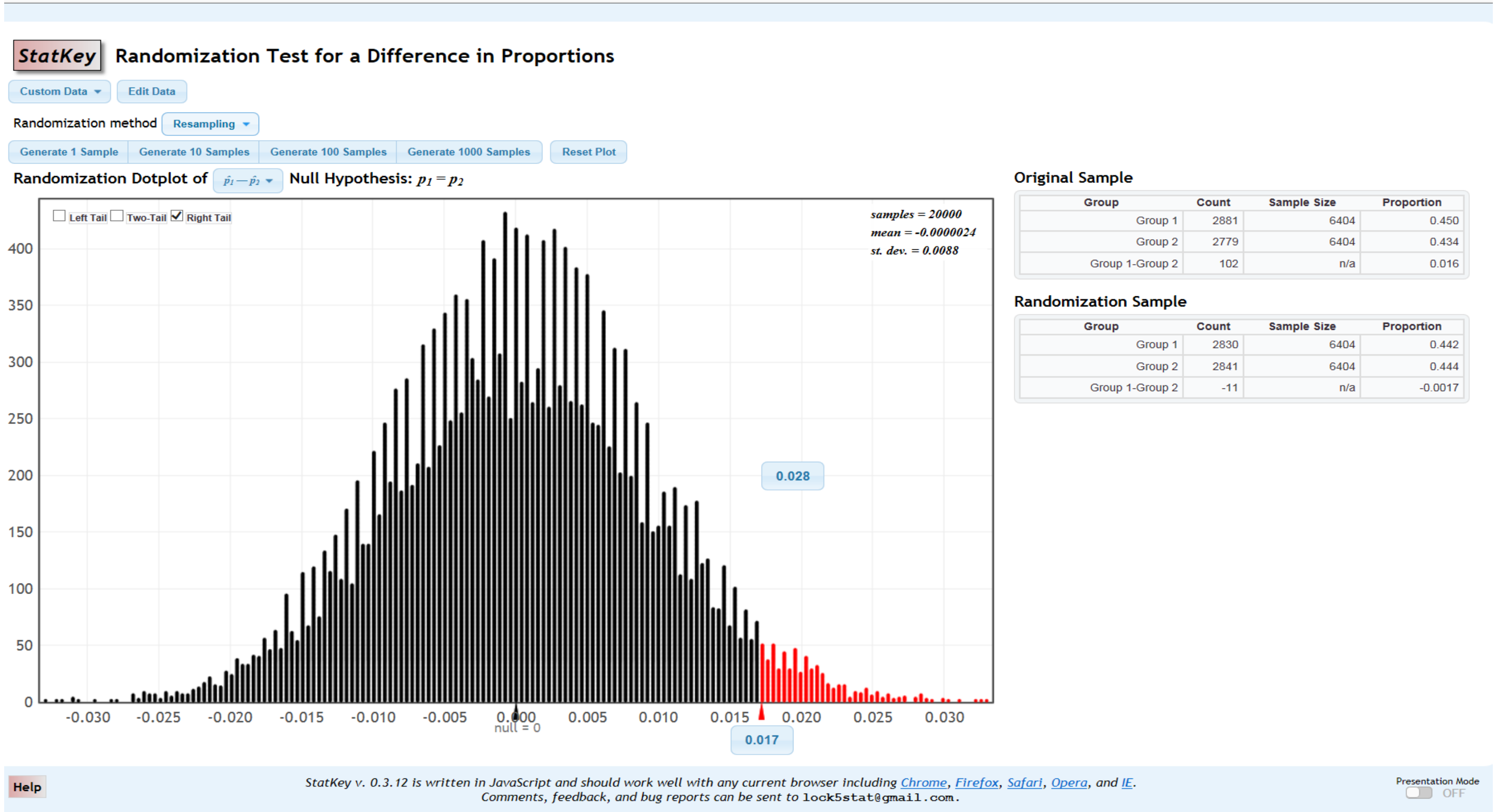
12.1. Wôhwɛâ akwadaa no mogya mu hwɛâ sɛ ôwɔ "HIV/AIDS"?		1. Yes	2. No	8. NK	HIVTEST
12.2. Wôhwɛɛ akwadaa no mogya mu ka kyerɛɛ wo sɛ w'anya "HIV/AIDS"?	1. Yes	2. No	8. NK	9. NA	HIVPOS
12.3. Apɔmuden dwumayɔni bi ka kyerɛ wo wô ayaresabea sɛ ɔdwene sɛ wo ba no anya "HIV/AIDS"?		1. Yes	2. No	8. NK	HIVHW
12.4. Obi wɔ abusua yi mu a dɔkota ayâ nhwehwâmu ahunu sɛ ɔbɔ nsaman wa?		1. Yes	2. No	8. NK	FAMTB
12.5. Na saa nnipa no ne saa akwaadaa yi te fie baako mu?	1. Yes	2. No	8. NK	9. NA	HOUSETB

13. INTERVIEWER COMMENTS AND OBSERVATIONS

Please write any additional comments or observations that you may have in this space.

END OF INFANT VPM FORM. THANK RESPONDENT(S) AND CHECK YOUR FORM.

Appendix C – Screen Shot of Statistical Significance Testing Using StatKey Software

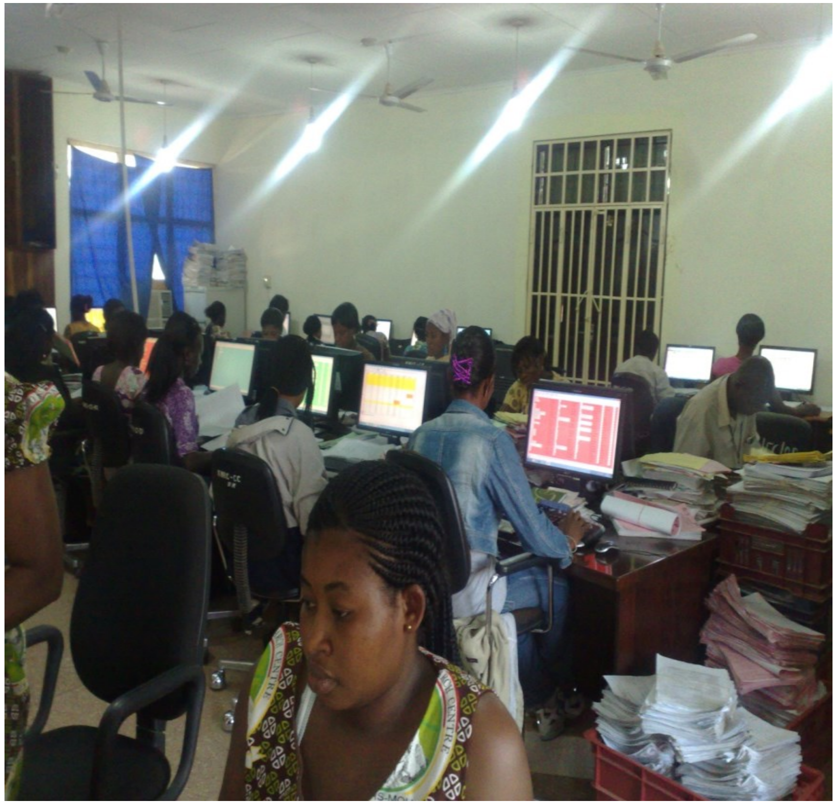


Appendix D- Transcription Work in Pictures

These are pictures showing the transcription of Verbal Autopsy documents carried out by the staff of the computer department at the Kintampo Health Research Centre in Ghana between November, 2011 and May, 2013

The author of this thesis (first on the left) Introducing the research to staff of computer department

Data entry clerk attempting to use the software.



Computer department staff entering the transcription of narrative.