HealthCare Tagging of Verbal Autopsies using SNOMED-CT

Rebecca West

MSc Computing & Management

Session 2009/2010

The candidate confirms that the work submitted is their own and the appropriate credit has been given where reference has been made to the work of others.

I understand that failure to attribute material which is obtained from another source may be considered as plagiarism.

(Signature of student)………………………………………………

Abbreviations

CoD: **C**ause **o**f **D**eath.

CSMF: Cause **S**pecific **M**ortality **F**raction. The proportion of deaths due to a specific cause.

CRISP-DM: CRoss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining is the industry standard methodology for data mining and predictive analytics.

EHR: Electronic **H**ealth **R**ecord

GATE: General **A**rchitecture for **T**ext **E**ngineering.

ICD-10: International **S**tatistical **C**lassification of **D**iseases and Related Health Problems –10th Revision is a coding of diseases and signs, symptoms, and external causes of injury or diseases.

IHTSDO: International **H**ealth **T**erminology **S**tandards **D**evelopment **O**rganisation. Owns and administers the rights to SNOMED-CT and other health terminologies and related standards

PAS: Patient **A**dministration **S**ystem

PCVA – Physician **C**oded **V**erbal **A**utopsy

SNOMED-CT: (Systematized **No**menclature of **M**edicine **- C**linical **T**erms), is a comprehensive health terminology that is used to code, retrieve, and analyze health data.

UMLS: Unified **M**edical **L**anguage **S**ystem

VA: Verbal **A**utopsy

VA Tool: Three components of a VA – questionnaire, mortality classification system and diagnostic criteria.

WEKA: Waikato **E**nvironment for **K**nowledge **A**nalysis **M**achine learning software.

WHO: World **H**ealth **O**rganization

WHO-FIC: World **H**ealth **O**rganization **F**amily of **I**nternational **C**lassifications

Technical Terms

Bayesian Analysis: A statistical technique for analyzing txt. Infers topicality from patterns of words and phrases present in documents. It is a “probabilistic method” because it returns a likelihood of a document belonging to a topic**.**

Class: A number of individuals (persons or things) possessing common attributes that are grouped together under a general or “class” name.

Classification: The systematic grouping of like things or objects into classes or categories according to some shared quality or characteristic

Corpus: a large and structured set of texts.

Feature: grammatical feature e.g. as the part of speech: number (single/plural) or gender assigned to a word.

Gazetteer: is a geographical dictionary or directory.

Gold Standard: is a diagnostic test or benchmark that is regarded as definitive.

POS Tagger: part-of-speech. Marking up the words in a text as corresponding to a particular part of speech, based on both its definition and its context.

Token: any word or other feature of a sentence that has a part of speech tag assigned to it.

Tokenizer: the operation of splitting a string of characters into a set of tokens

Concept: A concept is a clinical meaning identified by a unique numeric identifier (ConceptID)

Term: These can represent the terms that are in everyday use. There are often many synonymous descriptions for a single concept.

Sensitivity: The proportion of people with a disease who are correctly diagnosed (test positive based on diagnostic criteria). The higher the sensitivity of a test or diagnostic criteria, the lower the rate of 'false negatives,' people who have a disease but are not identified through the test.

Specificity: A statistical measure of how well a classification test correctly identifies the negative cases, or those cases that do not meet the condition. E.g. a medical test that determines if a person has a certain disease, the specificity of the test to the disease is the probability that the test indicates `negative' if the person does not have the disease.

Acknowledgements

I would like to thank the following people:

Dr Eric Atwell, my project supervisor for his guidance, support and insight.

Dr Katja Markert, my project assessor for her feedback.

Saman Hina who has helped me not only through her knowledge but also through her encouraging words throughout the duration of this project. I will always be grateful for her help and support.

Those who provided me with verbal autopsies samples Dr Karen Edmond at the London School of Hygiene and Tropical Medicine, London and Dr. Abraham D Flaxman and Sean T Green at the Institute of Health Metrics and Evaluation, Washington University, USA. i2b2 (Informatics for Integrating Biology and the Bedside), Boston, USA for the discharge summary data set.

My husband Martin, who supported me when I decided to leave work and take a new direction in my career, this course being the foundation. "It's been a roller coaster but together we have come out the other side!"

My sister Kirsty, although 12,000 miles away is with me always.

And finally to my Mum and Dad, what can I say? Always there, cheering me on through the good and the challenging times in my pursuit to achieve my hopes and dreams.

Summary

Verbal autopsy (VA) is widely used as a method of ascertaining cause of death in countries with incomplete or no vital registration systems. At present much VA interpretation is undertaken by physicians (physician coded known as PCVA) but this approach is resource hungry, expensive and can be inconsistent. Therefore, more cost effective alternatives need to be examined for assigning causes of death from VA.

There is significant interest in computers being able “assume” the role of the both the “coder” and “physician” to ascertain cause of death, although many challenges need to be addressed for this to become reality.

Although there has been much research into the subject of VA, most has been conducted in the epidemiological field. However, this report offers a systematic analysis from a computer science perspective and finds that the formal description and modelling of the problem space is fractured and poorly understood.

This project explores this issue and endeavors to describe, analyse and document the computational modeling problems associated with the verbal autopsy process and the steps required to address if computational solutions are to progress. Chapter 1 provides an overall background to verbal autopsies, the terminological systems which support them and other associated medical text, current approaches in natural language processing in the medical domain and data mining software which can assist in the computational process. Chapter 2 outlines how the overall project was managed and describes the three data sets that were acquired; American Discharge Summaries from the i2b2 challenge and two verbal autopsy data sets; one provided by the London School of Hygiene and Tropical Medicine, UK and the other from the Institute of Health Metrics and Evaluation, University of Washington, USA. In Chapter 3 the issues of challenges of Verbal Autopsy are documented and discussed. To illustrate, a computational prototype was built using SNOMED-CT Concepts, a nomenclature, GATE (text engineering tool), Python (program language) and WEKA (machine learning). As part of the process a detailed description on how the three data sets were prepared is provided, the modelling process and the prototype build together with all the issues and successes documented. Chapter 4 provides an evaluation of the both the prototype results and the systems used. In the final Chapter, conclusions are provided with recommendations on further improvements and future research required in this area.

Contents

3. Implementation of Solution

Appendices

List of Figures

List of Tables

Chapter 1: Background

This project topic was originally suggested by Karen Edmond and Betty Kirkwood of the London School of Hygiene and Tropical Medicine, and Sammy Danso of the Kintampo Health Research Centre, Ghana. They have conducted research on Verbal Autopsies [1,2] and approached Dr Eric Atwell at Leeds University to look into computational analysis techniques for Verbal Autopsies. Dr Atwell posted this as an MSc Project and I took on the challenge.

1.1 Verbal Autopsy: A Definition

Over half of the world's deaths go undocumented as to the cause [3]. This is in itself a tragedy. However, this also brings wider issues for major resource for health care planning and prioritization.

Countries that cannot record the number of people who die or why they die cannot realize the full potential of their health systems [4]. Rapid improvement of vital registration systems in many countries, although desperately needed is unrealistic. It takes considerable time and investment for countries to implement a reliable registration system with medical certification of cause of death.

Whilst the developed world has physician death certificates and autopsy data as the basis for their public health reporting, those in the developing world have adopted an alternative approach to support the information needs of their health care systems. Many have adopted the method of "verbal autopsy" (VA) - interviewing the relatives or caregiver about the symptoms and circumstances of a death and then interpreting the interview material to arrive at cause(s) of death [5]. A cause of death may be assigned by physician review of the questionnaires or by an algorithm [6].

1.2 Verbal Autopsy: Historical Background

In 1956, Yves Biraud recommended the uses of information supplied by the relatives of a deceased person in an attempt to establish *"a community diagnosis of the cause of death"* [7]. The first simplified lists of causes of death for use in developing countries were published by the WHO in 1978 [8]. The term "verbal autopsy" was first proposed by A.A Kielman in 1983 in his book an *"Analysis of Morbidity and Mortality"* [9]. However it is the work of Garenne & Fontaine who are considered the founders of the VA technique through the development of a VA questionnaire used in studies in Senegal [10]. This technique has been adopted worldwide [11]. There are currently 36 Demographic Surveillance Sites [DSS] in 20 countries, the Sample Registration (SRS) sites in India and the Disease Surveillance Points (DSP) in China who regularly use VA. [12]. A map of all countries using VA can be seen in Appendix C.

1.3 Verbal Autopsy: The Tools

A standard VA tool (see Fig:1) consists of a VA questionnaire, cause of death classification system and diagnostic criteria (physician review, expert or data driven algorithm) [5]. The actual questionnaire itself contains 10-100 questions [see Appendix D for an example]. There are two different interview methods [13]. One uses an in-depth, open-ended history of the final illness asking the care giver to outline the events in their own words. This is a descriptive account which will then be read and coded. The other technique is interviewer asking closed questions often pre-coded for use with an algorithm. Most VA's are conducted using a mixture of the both the closed and open-ended approach [13].

The interview is conducted by a well trained lay person, medically trained interviewer or health professional [14]. Much debate has taken place on the pros and cons of using lay and medical trained personnel. Although to date, the effects and outcomes of different interviewers are not known to have been formally studied [12]. Those conducting the interviews do receive training, although it is argued that the process would benefit from standardised guidelines. The understanding of local customs/culture, terminology and concepts of illness and their symptoms are seen as key in the process of acquiring a quality questionnaire [12]. The most common interpretation method of the questionnaire is local physician review without algorithms [6,15,16]. When the VA questionnaire is complete it is sent to a local health facility. On arrival the VA is annotated using the ICD-10 coding standards by a "coder" and then entered onto a computerized system either by the coder or a data entry clerk.

In this case each received questionnaire is reviewed independently by at least two physicians; when there is disagreement a third physician is brought in to review. If consensus can be gained a cause of death is decreed. If not, the death is recorded as "indeterminate". The second approach is expert algorithm. "*The algorithm can be developed from textbook description, existing clinical algorithms, local experience of a combination of both*" [15]. The third approach is data driven algorithm [17]. In this case each received questionnaire is reviewed independently by at least two physicians; when there is disagreement a third physician is brought in to review.

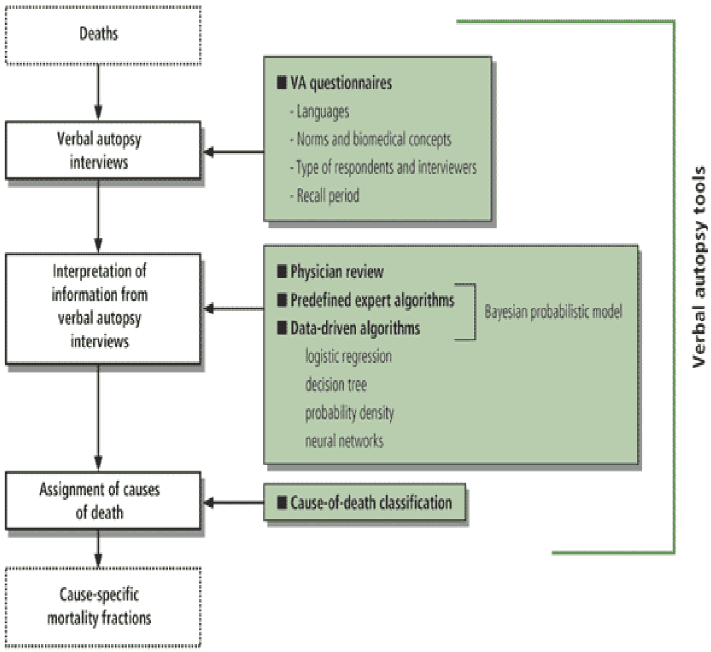


Fig 1.1 Verbal Autopsy Tools and Process. Source: Soleman et al 2006

If consensus can be gained a cause of death is decreed. If not, the death is recorded as “indeterminate”. The second approach is expert algorithm. “*The algorithm can be developed from textbook description, existing clinical algorithms, local experience of a combination of both*” [12]. The third approach is data driven algorithm [17]. This requires an additional sample of deaths from a medical facility where each cause is known and symptoms are collected from relatives. Then a parametric statistical classification method (logistic regression, neural networks and support vector machines) is trained on the hospital data and used to predict each cause of death in the community [14].

1.4 Terminology Systems

Another important facet to medical reporting and coding are the terminological systems which support the process. In its basic definition a terminological system is a system which contains standard terms denoting concepts and their relations which facilitate standardisation and control when recording medical data [18].

The subject of terminology systems is a challenging one. Literature on the subject was found to be unclear and often difficult to understand which was surprising considering the maturity of the systems and also that there are two organisations, International Standards Organisation (ISO) and Comite European de Normalisation (CEN), whose role is to clarify the standards [19]. However, the work of Keizer et al and Lusignan has very much helped to demystify them and to highlight the key characteristics of these systems, their purpose and their benefits [18,20]. For the basis of this project the term “terminological system” is an umbrella for the terms of “classification”, “thesaurus”, “vocabulary”, “nomenclature” and “ontology.” A terminology, thesaurus, vocabulary, nomenclature, or classification is called a coding system when the system uses codes for designating concepts.

To explain further; “terminology” is a list of terms, a “thesaurus” is ordered terms/synonyms, “vocabulary” are definitions, “classification” is a member of an arrangement, “nomenclature” is a composition of rules, an “ontology” is a set of concepts within a domain and the relationships between those concepts. Lastly a coding system is codes as designators [18-22]. Table 1.1 outlines the characteristics of the most well known.

Type	ICD-10	SNOMED-CT	UMLS
Terminology	**	**	**
Thesaurus	**	*	**
Vocabulary	NO	*	*
Nomenclature	*	*	
Classification	**	**	*
Ontology	NO	NO	*
Coding Schema	Significant	Significant	Non-Significant
	Hierarchical	Hierarchical	-
	Mnemonic	Mnemonic	Partly Mnemonic
	-	Juxtaposed	-

Table 1.1: Characteristics of Terminological Systems. Source Keizer 2000
**Acceptable for classification * partially acceptable for classification

Moving onto coding systems it is recognised that there are three generations coding systems [21]. First-generation is fixed organisation systems, e.g. ICD are typically hierarchical with simple structure such as a systematic list that is alphabetically indexed. Second-generation SNOMED-INT dynamic organization (i.e. provide multiple hierarchies) compositional, combining the simple list representation of concepts with a knowledge base to define and extend these concepts Third-generation systems e.g. SNOMED-CT, are based on formal models providing symbols denoting concepts and a set of formal rules to manipulate them

For terminological systems that have “significant” coding schema their structures are mnemonic, juxtaposition, hierarchical or a mixture of these. Mnemonic is when one or more of the characteristics is related to its class e.g. M = Male. Juxtaposition is when there are composite codes considering of segments which relate to the class e.g. in SNOMED-CT each medical concept has an individual concept id and from this terms (preferred terms and synonyms) and the relationships each with their own code are provided. Finally, there are hierarchical coding schemas e.g. in ICD10; “Endocrine nutritional and metabolic diseases” are E00-E90. Within this the “disorders of the thyroid gland” are E00-E07. Non-significant or “context free” coding schema have random or sequential coding.

Worldwide there are a number of terminology systems. This section covers the most significant systems and whilst outlining their purpose and functionality seeks to explore the interfaces and connections between them and their relevance and contribution to worldwide health care.

1.4.1 ICD Classification System

ICD is discussed at length in 3.1.5 explaining its merits the challenges associated to verbal autopsy. To provide wider context, ICD is used for morbidity and mortality statistics, reimbursement systems and automated decision support. The purpose of ICD is to increase international comparability in the collection, processing, classification, and presentation of these statistics

This classification has its origins in the 1850s. The first edition, known as the International List of Causes of Death, was adopted by the International Statistical Institute in 1893 [23]. The WHO took over the responsibility for the ICD at its creation in 1948 when it became the sixth revision. The classification system is regularly reviewed; minor updates are carried out annually with three-yearly major updates. It is currently in its tenth revision with ICD-11 planned for 2015 [24].

ICD is a core classification of the WHO Family of International Classifications (WHO-FIC). ICD is currently used in 193 countries and is available in the six official languages (Arabic, Chinese, English, French, Russian and Spanish) as well as being translated in 36 other languages. Twenty-five within the 193 countries use ICD-10 for reimbursement and resource allocation in their health care system [24-25].

The ICD-10 codes are broken down into 22 “chapters” with each chapter starting off with “Diseases of… [25]. ICD-10 codes consist of a single letter followed by 3 or more digits, with a decimal point between the second and third e.g. I21.0 "Acute transmural myocardial infarction of anterior wall**”**. A full list of the ICD-

10 chapters can be found in Appendix E. Arguably, ICD in terms of coverage, impact and usage is the most influential and important health classification system in the world.

1.4.2 SNOMED-CT

SNOMED-CT was created in 1999 through a joint development between the National Health Service (NHS) in the UK and the College of American Pathologists (CAP). The international clinical terminology was created by the convergence of SNOMED-RT and the UK's Clinical Terms Version 3 [26-28]. In 2007 management of SNOMED-CT was transferred to the International Health Terminology Standards Development Organisation (IHTSDO), a not-for-profit-making organisation based in Denmark.

SNOMED-CT is considered to be the most comprehensive multilingual health terminology in the world [26-28]; achieved through the development of a built-in framework to manage different languages and dialects. SNOMED-CT is available in English (both UK and US), Spanish and Danish with translations into Swedish, French and Lithuanian. There are plans to expand the translation of the standard into other languages.

SNOMED-CT has more than 400,000 unique concepts [26]. The concepts are organized in hierarchies enabling very detailed clinical data to be recorded, accessed or aggregated. Each concept is represented by an individual number. The example below shows how SNOMED_CT represents "Myocardial Infarction". What lay people would refer to as a "heart attack". In SNOMED_CT, Myocardial Infarction has the Concept Id: 22298006. SNOMED-CT also states the preferred term and synonyms associated with this disorder and if appropriate any homonyms.

In this case the preferred term is "Myocardial Infarction". The synonyms being "infarction of heart", "heart attack", "MI", "cardiac infarction" and "myocardial infarct". There are no homonyms in the example.

SNOMED-CT has the ability to cross map codes from the legacy systems: "Myocardial Infarction" and also lists both the SNOMED-RT id: in this case it would D3-1500 and with the clinical terms code CTv3 id: X200E. An example structure of SNOMED-CT concept see Appendix F.

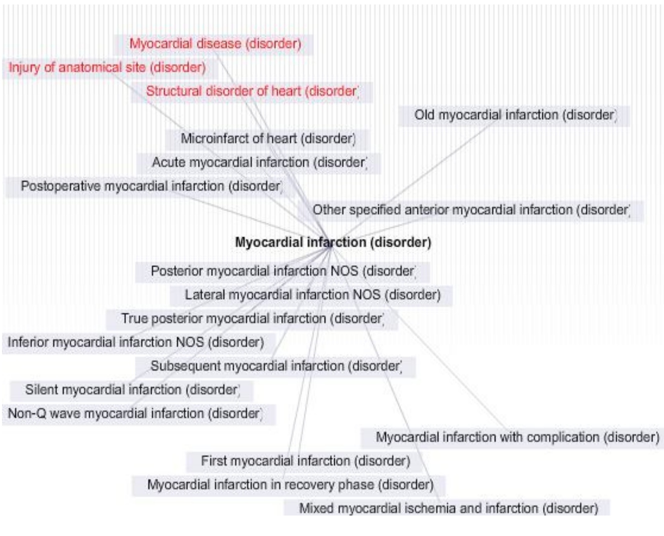


Figure 1.2: SNOMED-CT Example: Myocardial Infarction

The referencing of conditions and symptoms using individual numbers provides a number of benefits: the elimination in confusion of local terminology and the standardisation of language which supports the exchange of clinical information. Therefore, SNOMED-CT aims to provide consistency and interoperability through the standardisation of medical terminology. In terms of its impact, again it is significant, with SNOMED-CT used in over 50 countries and growing [26].

1.4.3 UMLS

The Unified Medical Language System (UMLS) was created in 1986 by the US National Library of Medicine [29]. It is a database of numerous biomedical science vocabularies. It contains a mapping structure against these vocabularies enabling translation among the various terminology systems. It is also considered a comprehensive thesaurus and ontology of biomedical concepts. In this respect it has similarities to SNOMED-CT. However, UMLS has the addition of a lexicon which is used for natural language processing used mainly by developers of systems in medical informatics. The UMLS is composed of three “knowledge sources”; Metathesaurus, Semantic Network, Specialist Lexicon [30].

1.4.3.1 The Metathesaurus

The Metathesaurus contains 1 million biomedical concepts and 5 million concept names, in 17 languages sourced from 120 incorporated controlled vocabularies and classification systems which include ICD-10, SNOMED-CT in 17 languages [29-41]. The Metathesaurus is produced by the automated processing of machine-readable versions of the source vocabularies, followed by human intervention of editing and review. It is distributed as an SQL relational database and can also be accessed via a Java object-oriented API [29-31].

1.4.3.2 Semantic Network

Each concept in the Metathesaurus is assigned to at least one "semantic type" (a category), and certain "semantic relationships" may occur between members of the various semantic types. The semantic network is a catalog of these types and relationships. Currently there are 135 semantic types and 54 relationships [29-31].

1.4.3.3 SPECIALIST Lexicon

The SPECIALIST Lexicon contains information about common English vocabulary, biomedical terms, terms found in MEDLINE and in the UMLS Metathesaurus [29-31]. Each entry contains syntactic, morphological and orthographic information. A set of Java programs use the lexicon to work through the variations in biomedical texts by relating words by their parts of speech, which can be helpful in web searches or searches through an electronic medical record.

Finally, UMLS has a number of supporting software tools, one of which is MetaMap, an online tool which when given a piece of text, finds and returns the relevant Metathesaurus concepts.

1.4.4 Critical Evaluation of Technological Systems.

There have been many research papers evaluating the performance and making comparisons between the different technological systems [32-36]. The reviews paint a mixed picture and no overall agreement has been gained. This is not surprising. To explain, it is difficult to compare the utility of different coding systems. History is important. To illustrate, it is important to recognise the origins of the terminologies; SNOMED-CT had its origins in both pathology and primary health care through its connections with CAP and the NHS. ICD's roots are in mortality and morbidity. UMLS is a collection of many vocabularies. Although over time it could be argued that all three have evolved to become more general purpose terminology systems. Therefore, it is not surprising that in various studies when comparing the ability of different systems to code patient records that SNOMED-CT and UMLS (which contains SNOMED-CT) outperforms ICD-10 in this area [35,37].

Ultimately, whatever system is used its perceived merits and potential shortcomings depend on the purpose of the system, how it is being used and whether it meets and satisfies the needs of the user, whether that is an individual, organisation, health provider etc.

What can be agreed as "common ground" is that all the systems seek to standardized clinical terminology to enable machine readable clinical data to aid the reconciliation of the representations made when using natural language.

In relation to this project there were a number of reasons why SNOMED-CT was chosen. Nomenclatures are the most sophisticated of all the terminologies allowing concepts to be combined to enable more complex concepts to be created [18,20]. As a direct result it has finer concept granularity and a richer expressiveness. As the source data for the project were verbal autopsies and discharge summaries, both of which contain a significant amount of free text, it was deemed that SNOMED-CT would have some possible advantages over other classification systems [18]. The wide range of concepts and ability for composite use provides abstract data extraction rather than single terms. Although it was recognized that nomenclatures are significantly larger than classification systems, therefore much more complex and could provide issues at data preparation and deployment stages. Finally, availability of terminological systems was another consideration. To obtain access to ICD-10 or UMLS licences would need to have been sought which would have taken time and also there were no guarantees that these would have been granted. Full access to the SNOMED-CT was granted through undertaking some support work for the NIH National Center for Biomedical Computing i2b2 informatics for integrating Biology and the Bedside [38] Challenges in Natural Language Processing for Clinical Data. NB: "Deidentified clinical records used in this research were provided by the i2b2 National Center for Biomedical Computing funded by U54LM008748 and were originally prepared for the Shared Tasks for Challenges in NLP for Clinical Data organized by Dr. Ozlem Uzuner, i2b2 and SUNY".

1.5 Natural Language Approaches to Medical Text Analysis

Natural language approaches have evolved to encode medical data. At first the NLP technologies only parsed the data and were unable to encode them using terminological systems [39]. The Symtxt system, the statistical NLP tool, MEDSYNDIKATE, Genia Tagger/Genia Corpus and MEDIE and MetaMap/MMtx are all examples of medical NLP systems [40-45]. Historically, a key challenge with medical NLP tools has been that they have not been easy to adapt or reuse. One reason is that medical NLP programs are often tailored to domain or institution-specific document formats.

However the development of MedLEE by Carol Friedman in 1995 revolutionized this area of research and it became one of the first NLP technologies to perform consistently and effectively in extracting clinical data through the use of clinical ontologies [46-51]. MEDLEE was launched into the commercial domain in 2008 [52].

1.5.1 MedLEE

The Medical Language Extraction and Encoding System (MedLEE) is a natural language processor that identifies clinical information in narrative reports and maps them to a controlled vocabulary [47]. When first developed MedLEE mapped radiology terms to the Medical Entities Dictionary (MED). However, the system now maps UMLS concepts based on structural matching using modifiers [47]. MedLEE uses lexical and semantic rules to regularise terms identified in documents. A regularised term is looked up in the UMLS knowledge source and suitable UMLS concept identifiers are returned as matches. Below is an example showing the academic version of MedLEE. The document to the left is a sample discharge summary in its pre-processed state, then to the right is the output from MedLEE once the clinical concepts have been extracted and tagged.

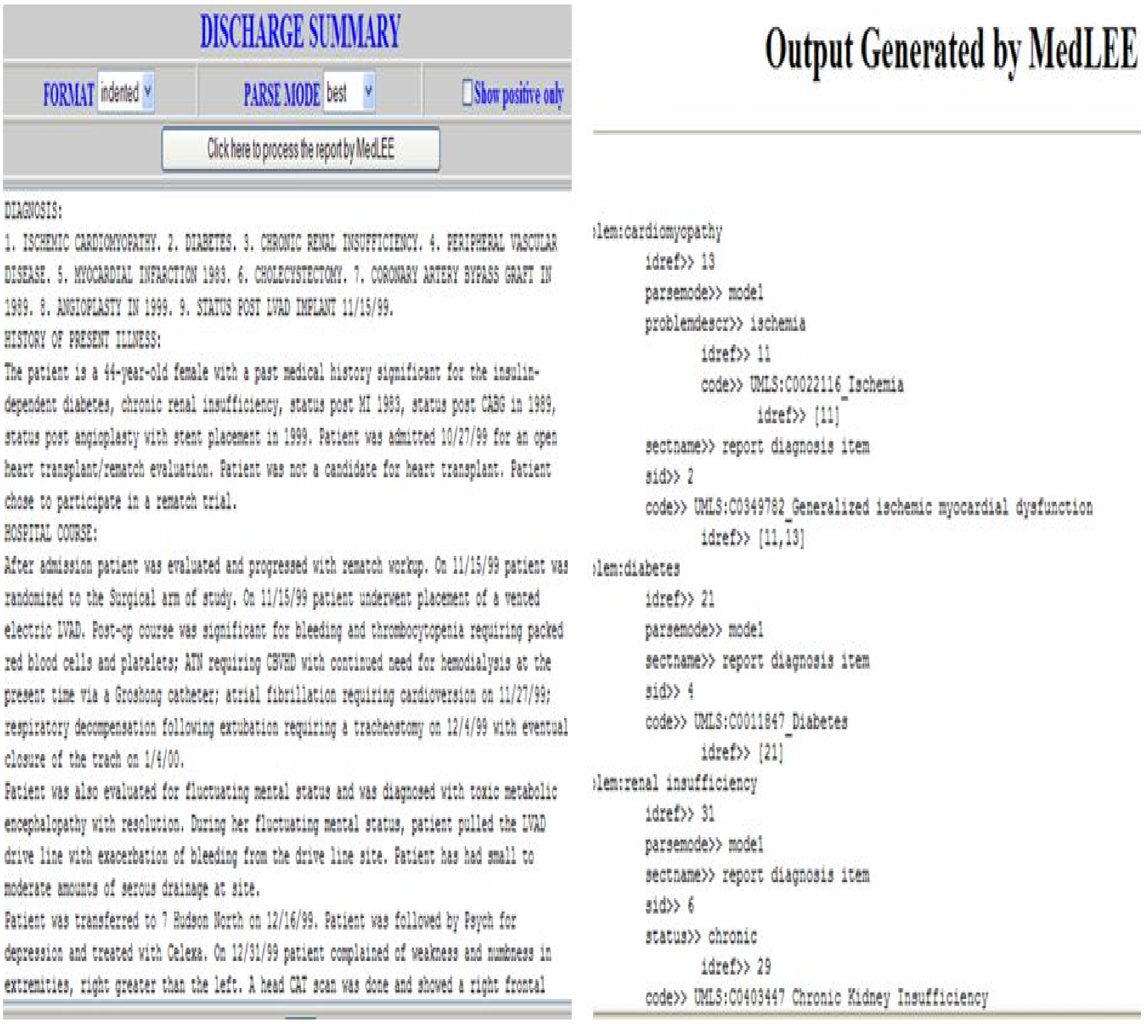


Figure 1.3: Discharge summary containing clinical concepts (left) are extracted and tagged to UMLS concepts, output shown (right) (Source: MedLEE website).

Since MedLEE there has been significant amount of research in lexicon-semantic mapping of various medical terminologies to the UMLS and other terminologies [40,50,53-57].

However, this area of research needs to be continued in development as there is still much to do to build new NLP systems to advance the capabilities for mining and coding clinical text. One of the most influential key research "hubs" in this field is The NIH National Center for Biomedical Computing "Informatics for Integrating Biology & the Bedside" (I2B2), whose purpose is to encourage learning and the development and distributing of open source software for NLP in clinical records [38]. This group aims to drive the research forward bringing together medical informaticians, natural language researchers, processing researchers and data owners. The clinical challenge is now in its fourth year.

In terms of this project the information from i2b2 provided a number of benefits – access to the SNOMED-CT nomenclature but also shared learning on the available tools and developments. This enabled a comprehensive list of NLP resources to medical text analysis and extraction to be built, see Table 1.2 with a more detailed description in Appendix G. However, the greatest benefit was reading about one of the NLP tools used 2006 Challenge, a NLP tool developed called the Health Information Text Extraction (HITEx) tool [58]. What was particularly interesting was how GATE (General Architecture for Text Engineering) could assist in the annotation of clinical terms. This led to onward reading into GATE where it was established that it was open source software that had an ability to process a wide range of text. As a result this was selected as the text engineering tool for the project.

	SOFTWARE		SOFTWARE
2	Berkley Parser	18	MEDSYNDIKATE
4	BIOSimply	20	Meta Map
5	CCG Parser	21	MOBY
6	ClearTK	22	Natural Language Toolkit
9	dTagger	23	NegEx/ConText
10	ENJU	24	OpenNLP
11	GATE	25	Python
12	Genia Tagger	26	SimFind
13	MALLET	29	Stanford Parser
14	MedEx	30	SYNTXT
15	MEDIE	31	UCLA Medical Imaging Informatics Toolkit
16	MedLEE		
	OTHER RESOURCES		OTHER RESOURCES
1	Banner	19	MeSH vocabularies
3	Bioscope Corpus	27	SNOMED-CT
7	cTakes	28	Specialist Lexicon
8	DrugBank	32	UMLS vocabularies
17	MedRA	33	WordNet

Table 1.2: List of NLP Software and Other Resources

1.5.2 GATE

GATE has been developed by the University of Sheffield. GATE is an open source text analytics software tool which is able to process a wide range of text data [58-60].

GATE is an architecture, a framework and a development environment for Language Engineering. [59]. GATE is a component based model with the components being one of three types of Language Resources; (LRs) represent lexicons, corpora or ontologies, Processing Resources (PRs), which contain common NLP tasks e.g. tokeniser, part-of-speech (POS) tagger, gazetteer etc. These processing resources grouped together are known as “ANNIE” in GATE “A Nearly-New IE system. Lastly, there are Visual Resources (VR’s) which enable visualisation and editing of components within the GUI [58-60].

For this project to enable the prototype to be built successfully, GATE was used to build a semantic annotation pipeline including all the appropriate “rules” to enable optimum performance and the development of a new gazetteer using the source SNOMED-CT concept file. The acquired medical text (discharge summaries and verbal autopsy) were pre-processed and then loaded into GATE to form the corpus. The annotation pipeline was run over the corpus to “tag” the medical concepts for each document. In relation to VA, in its most simplistic description the “GATE process” was in place to attempt to fill the role of “VA coder”. The aim, to assess its competency at term identification and coding, drawing out any computational/NLP issues. The results were then passed to a classifier to determine if an accurate cause of death could be determined.

1.6. Machine Learning Software/Data Mining Software

Although there are a number of machines learning tools/software available, e.g. RapidMiner and ELKI [61-62], WEKA was the chosen tool to build the classifiers. WEKA was chosen primarily as it was a known entity, currently used at the University but also it is well established and well regarded both in academia and the commercial arena across the world [63-64]. Finally, it supports process models of data mining including CRISP-DM which is the chosen methodology for this project [65-66].

1.6.1 WEKA

Developed at the University of Waikato, it has a comprehensive collection of machine learning algorithms which include regression, classification clustering, and data preprocessing tools [64].

For the project, once the data has been extracted and annotated via the GATE tool, the data will then be prepared converting into an ARFF file to run on a classifier and to return some meaningful results for evaluation. The results aim to understand the issues and successes of using data driven algorithms and to understand how effective a computational approach would be to replace the physician's decision judgment in ascertaining cause of death.

1.7 Additional Support for Prototype: Use of Python

Once the project was underway at the prototyping stage it was established that GATE was unable to output the annotated medical concept terms. Since a format of CSV or ARFF was the required input for WEKA, a python program was written to process the annotated medical concepts once the document had been passed through GATE and the frequency of the word occurrence captured and then output into SV or ARFF format. This enabled the process to remain an automated one rather than having to move to manual recording of GATE output.

Chapter 2: Design of Solution

2.1 Business Understanding: Project Planning and Management

To enable successful project content and delivery, weekly project meetings took place with the project supervisor since March 2010. A project plan and a blog site (http://mscgirl.wordpress.com/) were built, regularly input to and reviewed. Both served to track progress against key milestones actions and facilitate discussion. A copy of project plan can be found in Appendix H. A presentation was also prepared and delivered at the progress meeting in July, see Appendix I.

2.2 Working to the Project Plan

In terms of working and keeping to the project plan, all milestones were on schedule at the point of the interim report production bar one, the acquisition of a verbal autopsy sample**.** Up until then only the discharge summaries were available**.** At this point a new approach had to be taken to by continuing to use the discharge summaries to build the prototypes and gain learning and knowledge on the process. When the verbal autopsy samples arrived it was found that there were some similarities between the documents but there would be some additional challenges. These are written in detail in the subsequent chapters, although in essence it meant some python programming had to be injected into the project, some changes within the prototype phases and a week of the two weeks contingency time built in at the start of the project had to be used.

2.3 Literature Review

The initial concern was the likelihood of high volume of academic research papers and sources of information. The initial literature search acquired a number of seed papers. Keywords searches using Google scholar and PubMed for "verbal autopsy", "verbal autopsies", "discharge summaries", "NLP and clinical text", "Data mining and clinical text". Through forward and backward reading three target strands emerged: medical text sources, terminological systems and data mining (see Fig 2.1). Other valuable sources of information came from research groups in the medical text analytics; predominately University of Sheffield (NLP Group). All the sources of knowledge were reviewed to ascertain overlaps and then conjoined together. In total over 240 research papers were reviewed. Although over 140 were discarded as they were either too steeped in medical influence or provided similar content. From this the project took shape and the scope became clear.

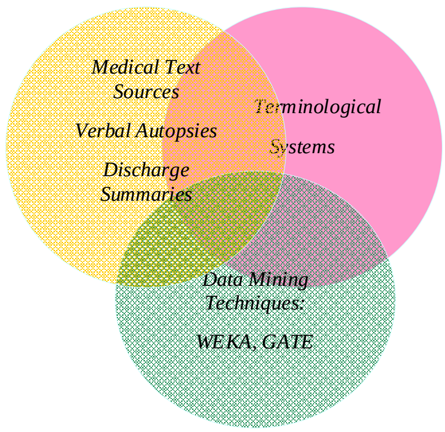


Figure 2.1:
The Three Target Strand Approach

2.4 Project Methodology

In terms of research project methodology the CRISP-DM (Cross Industry Standard Process for Data Mining) Process) model was used [66]. The rationale being that it is an excellent fit to this project. To explain, on commencement of the project although the subject area had been identified, the requirements and aims of the project were fluid and flexible. To obtain the best outcome it was crucial to build and refine as knowledge and experience grew. As a result, without firm and exacting requirements the waterfall methodology was rejected. The spiral methodology was another consideration although at the time of project commencement due to the steep learning curve required it was felt that tackling the most difficult aspect of the project without a full grasp of the background material would only prove to be a more lengthy process in the long term and likely to less support the delivery of the project.

So in conclusion, thought was given to this fundamental question - What type of project is this? In essence it's about understanding a problem space and through this building a prototype with a number of iterations to understand the issue and draw conclusions. Thus the projects core is text analytics and data mining. In view of this the CRISP-DM was considered to be best fit. The model comprises of six stages: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment [66] (see Fig 2.2).

The outer circle symbolizes the cyclic nature of data mining. The outcome of each phase determines which phase or task within a phase to be performed next. The arrows indicate the most important and frequent dependencies. The data mining process continues after a solution has been deployed. The lessons learned during the process trigger new ideas or questions. In following this model, subsequent data mining processes will benefit from the experiences of previous ones. For this project all stages will be conducted, except the deployment stage, usually this is the stage of business launch instead it is the production of this report. Overall this approach is an excellent fit to this project. Why? Strong emphasis needed to be placed on a thorough understanding of the dataset and its preparation

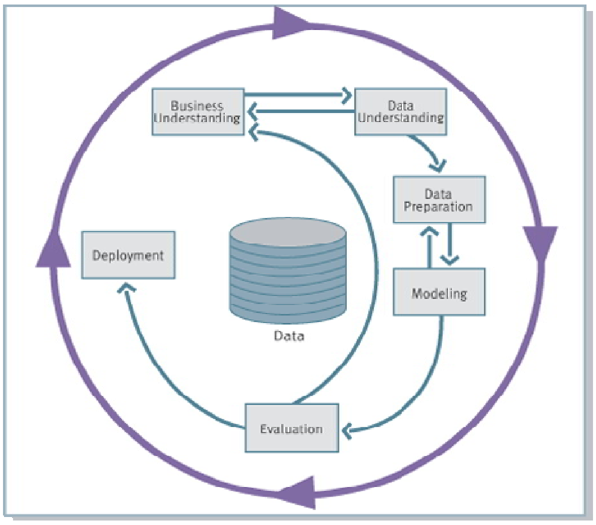


Figure 2.2 – Phases of the CRISP-DM Process Model
(Source: http://www.crisp-dm.org/Process/index.htm)

This was crucial to enable the accurate extraction of the medical text terms and the mapping using SNOMED-CT. The model supports this approach. The project required an iterative approach again which this model supports with the flexibility to move back and forth between the phases. The modelling stage appeared to hide a significant amount of work, this was the prototyping stages which needed to be to revisited on a number of occasions and then evaluated. It is anticipated that that the application of this methodology will prove invaluable to the achievement of the project aims and objectives.

2.5 Project Aims

The project aimed to fulfill both the minimum and additional requirements, both of which are detailed below.This project has a two pronged approach; its aims describe, analyse and document the computational problems associated with the verbal autopsy process examining the steps required to address if computational solutions are to progress. To illustrate the challenge a computational prototype has been built. Although all the output results are detailed in this report and they are very important, so too is the understanding of the problem space and the recommendations on how further improvements and future research needs to develop.

2.5.1 Minimum Requirements

- To understand the purpose and value of verbal autopsies.
- To understand the current VA processes and any issues associated with these processes.
- To explore past and present academic research conducted on verbal autopsies and other medical text.
- To obtain a sample of English medical text data to perform text analytics extracting the key concepts using the SNOMED-CT codes and descriptors.
- To build a prototype automated tool for classification.

2.5.2 Additional Requirements

- Clean noisy data from the medical text to improve the overall quality of the tool and overall diagnostic results.
- Ability to extract the medical terms from both the structured and unstructured data.
- Evaluate the prototype and identify avenues for enhancement, with view to making improvements.

2.6 Data Understanding: Acquisition of the Data Set

From the outset of the project at least two medical data sets were required, a sample of verbal autopsy data and a file containing medical concepts.

At project commencement the medical concept [SNOMED-CT] file was available; however verbal autopsy data was not. So whilst working on establishing a source for this data, a sample of 350 discharge summaries from the USA was acquired. Discharge summaries are "*A clinical report by a physician or other health professional at the conclusion of a hospital stay or series of treatment. It outlines the patient's chief complaint the diagnostic findings, the therapy administered and the patient's response to it and recommendations on discharge" [67].* Discharge summaries were used as they have parallels to verbal autopsies in that there are both examples of medical text and both are steeped in natural language containing unstructured, ungrammatical and fragmented information [30,35,]. Through acquisition of the discharge summaries this enabled a full prototype to be built mirroring all stages within the verbal autopsy process.

2.6.1 Discharge Summary Sample

Through the author's links to the most recent i2b2 Challenge Informatics for Integrating Biology and the Bedside [38] it was possible to source 350 discharge summaries. To gain access to the discharge summaries it was mandatory to undertake and "pass" a 3 hour web based course set by the National Institute of Health to demonstrate a level of competence and knowledge on "Protecting Human Research Participants" togetherwith signing a data agreement. Both were completed and a copy of the certification can be found in Appendix J. The discharge summaries came from Partners HealthCare, 97 in total, Beth Israel Deaconess Medical Center, 73 in total and lastly 180 summaries from the University of Pittsburgh Medical Center [38]. An example of one of the discharge summaries can be found in Appendix K.

2.6.2 Ghana Verbal Autopsy Sample

There was an expectation that a considerable sample of verbal autopsies could be obtained through the London School of Hygiene and Tropical Medicine. Unfortunately, and disappointingly, this turned out to be not the case despite many requests for the data over the complete duration of the project, both from the student and indeed the project supervisor. In the end a sample of 5 was provided 22nd July 2010. The samples were all cases of neonate deaths (children from 0-28 days old). An example of one of the verbal autopsies can be found in Appendix D.

2.6.3 IHME Verbal Autopsy Sample

When it became apparent that there were going to be difficulties acquiring verbal autopsy data, then other avenues had to be sought. This proved to be an almost impossible task due to data protection issues (preventing the release of data) and a lack of contacts in the medical field that could assist with the acquisition. In discovering this registration was applied for and accepted to gain access to Measure Demographic and Health surveys (DHS) which are funded by US AID. The DHS projects purpose is the production of surveys to advance the global understanding of health and population trends in developing countries, which includes VA data. [68]. Unfortunately on being given access to the surveys the data was found to be un-processable due to the need to have access to commercial statistical analytics (SPSS, SAS or STATA). At this point early July real concerns were developing as to whether a sample could be obtained. Through doing some research on machine learning and verbal autopsies a research paper was found which had been produced by two individuals based at the Institute of Health Metrics and Evaluation (IHME) in Washington [69,70]. Through contacting these individuals a CSV file of data values derived from 1592 verbal autopsies was obtained. The sample is referred to as “IHME” throughout the project report.

2.6.4 SNOMED-CT Data File

Through links to the i2b2 challenge, the SNOMED-CT file was provided in a raw text document. The document was of significant size 28 meg and through assessing this needed to be cleaned to extract only the concepts from the file.

2.7 Description and Exploration of the Data

Initial views of the exploration of the data were that overall it was disparate in terms of size, content and format. Although there had been research papers written only using one example of medical text or very small samples and conclusions drawn from the experience [53,71] a larger data set would have been preferable. So considerable thought had to be given of how to work through the datasets to best effect to illustrate the computational techniques to support automated VA cause of death diagnoses.

2.7.1 Discharge Summary Sample

The summaries were provided as raw text. Initial observations on the data was that they were pre-processed in terms of having the personal health information (PHI) removed to ensure patient and physician anonymity. In terms of gold standard inclusion there was no separate file detailing the gold standard as it was contained in the “diagnoses” section within the summary. This was initially considered as an issue. However, on balance it was deemed relatively unimportant. To explain, consideration was given on whether to write a program to extract the final diagnosis from each discharge summary. It is possible to identify the diagnosis section within the discharge summaries as each section is usually with a label in upper case and separated with a colon, see fig 2.3. Through pattern labelling a program a computer can be taught to look for section names and hence read, identify and allow the extraction of the final diagnosis [71]. However, on review it was discounted as an exercise. There were two main reasons; one because when looking at the discharge summaries the terminology used for the diagnosis section of the discharge summaries varied. Some used FINAL DIAGNOSES, PRINCIPAL DISCHARGE DIAGNOSIS, and DISCHARGE DIAGNOSIS. Clearly this would add to the already computational effort required to build such as program. But secondly and more significantly the project was about verbal autopsies and ascertaining a correct cause of death. With VA documents there would be no extraction activity of the specific cause of death within the text, as this is absent as it is only cited on the death certificate. Although, what this did identify from a text analytics perspective is that just a simple task of getting a computer to understand the semantic meaning of a very simple heading such as “diagnoses” is extremely challenging and that it is compounded with the fragmentation between computer software systems which record and store this information.

FINAL DIAGNOSES :

1. Coronary artery disease.
2. Acute myocardial infarction.
3. Complete heart block, status post recent permanent pacemaker implant at **INSTITUTION

BRIEF CLINICAL HISTORY:

This is an **AGE [in 80s]- year - old male who initially presented for evaluation at **INSTITUTION where he was complaining of dizziness that had been going on for approximately 1 month He is a patient of Dr. **NAME[QQQ PPP] and carries a history of congestive heart failure that has been treated medically , also has had removal of a skin cancer from around the left eye .

Fig: 2.3. Excerpt from the US discharge summary sample.

2.7.2 Ghana Verbal Autopsy Sample

The Ghana verbal autopsy sample came in two formats; Format 1 consisted of five word documents with a separate file detailing the gold standard cause of death diagnosis (see Appendices D and L). Long in duration, each document was circa 18 pages and contained both very structured and free text formats. Clearly the documents were in a different format to the discharge summaries. Format 2 consisted of a CSV file detailing all the responses to the questionnaire in total. This detailed 246 attributes of one of the following data types categorical, binary and continuous. On checking all the symbolic fields ("yes" "no" don't know") had been set to numeric values. This was important to recognize as modeling tools/algorithms often require this format to enable processing. The verbal autopsies were all examples of neonate deaths; this raised concerns on how effective the SNOMED-CT concepts would be on annotating these documents considering the diseases and symptoms derived from a very distinct area of medicine and also seemed steeped in local terminology. The true impact would not be known until the results of the prototype were established.

2.7.3 IHME Verbal Autopsy Sample

The IHME verbal autopsy data acquired was in "csv" format. A gold standard cause of death diagnosis was included within the CSV file. Although unlike the Ghana sample the cause of death reason was heavily anonymised, just stating a code between 1 and 32 for cause of death. Although this in itself was not catastrophic with the sample what it did mean was that it would be difficult to evaluate the statement from the research findings documented in Chapter 3 (3.1.4) that stated that data driven algorithms found it harder to draw conclusions where there were no clear water between the symptoms of the disease. Also if the country of origin was known it would have added further context to the results which would have been useful. Similar to the Ghana csv file all the symbolic fields had been set to numeric.

2.7.4 SNOMED-CT Data File

On examining the SNOMED-CT concept file, a quick assessment of the raw text file showed that it indeed contained nearly 400,000 medical concepts. Opening the file it was clear that some clean up would be required. The concept file contained 6 sets of data CONCEPT ID, CONCEPT STATUS, FULLY SPECIFIED NAME, CTV3ID, SNOMED ID AND ISPRIMITIVE (see Appendix M). The only information that was required was the FULLY SPECIFIED NAME, the full and preferred term which was used in SNOMED coding. Also after each FULLY SPECIFIED NAMED it had a further annotation; for example myocardial infarction **disorder.** Within SNOMED-CT there is a top level hierarchy in which concepts are classed e.g. a disorder, finding, procedure, substance etc see fig 2.4. Any reference to these would need to be removed before the data could be loaded into GATE. This was achieved by writing a simple program in python to remove these references.

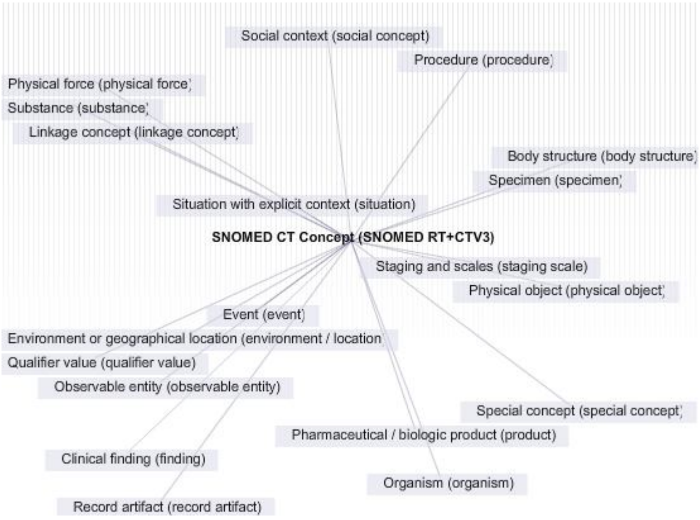


Fig 2.4 Hierarchies within SNOMED-CT

2.8 Data Quality

Data quality can be assessed in different ways. In terms of this particular data set, it would be fair to comment that overall the data quality was of an adequate standard. It is important to note here that if the data sample size was not included in the assessment then overall the data quality would have been considered to be good. However, it is sample size that has given the data an assessment of adequacy. The added dimension of having three disparate data sets rather than one also added complexity into the project at all stages.

2.8.1 Quality Evaluation of the Discharge Summary Sample

Overall a very comprehensive data sample, format is readable and very processable and the corresponding gold standards are within the documents. Country of origin is known and also there is an intimation of age group within the summaries which help with context. The data set came from three different sources, so there were some clear “style” differences noted within the free text and this would need to be observed at the modelling and evaluation to stage to see if this had any impact (positively or negatively) on the overall results. In terms of format, as previously noted, there were different section heading titles used within the summaries although this was not viewed as majorly significant; again this would be evaluated at the results stage. It was noted that within the 180 summaries from the University of Pittsburgh Medical Center, 81 were actually progress summaries whilst the patient was in the care of the hospital. These needed to be reviewed at data preparedness stage and evaluated for suitability and inclusion.

2.8.2 Quality Evaluation of the Ghana Verbal Autopsy Sample

On observation the sample had quality in terms of content both in terms of the document accompanied by the gold standard and the alternate CSV file but the lack of sample size was an issue. Also, unlike the discharge summaries and IHME sample, the questionnaire had questions about both the baby and the mother's health. This added an additional dimension; one could argue complication, to the text annotation and extraction and raised the question on how this could be addressed, if at all through the use of GATE. The document itself was very comprehensive which from one angle, if it was being observed with human eyes and experience would provide an excellent overview of the signs and symptoms to ascertain a cause of death. However with the sheer volume of data, coupled with the very structured approach and varying data types, there were concerns over the ability to output meaningful findings using a computational process. Also it was very evident that the document, especially in free text areas, had spelling mistakes of both medical and non medical words. Clearly that provided an authentic experience to run an experiment although this again raised concerns with the annotation process.

2.8.3 Quality Evaluation of the IHME Verbal Autopsy Sample

The sample was in csv format. When the file was first obtained no details where available to explain which attributes were of which data type, see Fig 2.5.

```
"symptom1","symptom2","symptom3","symptom4","symptom5",...,"causeOfDeath"
0,70,1,2,1,0,...,14
...
```

Fig: 2.5. Extract from IHME Verbal Autopsy Sample

All that could be established was that there were 32 causes of death and 142 "symptoms". Through email "persistence" (see Appendix N) it was established that the file contained a variety of data; categorical, binary and continuous. Through the email exchange it was also established that not all the 142 "symptoms" were actually symptoms i.e. an indicator of disorder or disease, in fact they were **all the** attributes which were contained within the questionnaire. Although this was helpful and provided more accurate results with the prototype it did have implications for the project. This information only came to light late July and as a result all the experiments with this data sample had to be completely redone. This raised the issue of ambiguity within data samples and the need to clearly express the contents to ensure that results and true and valid. This in itself highlights the disconnect between the understanding and the reporting of the data by various stakeholders in VA process. As a lay individual I experienced some confusion when reading the research articles of experiments conducted on the effectiveness of the various VA tools – PVCA, data driven and expert algorithms. Often it was unclear about the treatment of the sample, its shortcomings and the exact methods employed to ascertain and extract information from the sample.

2.8.4 Quality Evaluation of the SNOMED-CT Data File

In view that the SNOMED-CT supports a worldwide health care demand with an excellent reputation and track record the quality of the data was not in doubt. However, the most important quality aspect of the file was to ensure that the cleansing of the SNOMED file was done properly and no integrity issues were introduced into the data file.

Chapter 3: Implementation of Solution

3.1 Understanding the Issues and Challenges with Verbal Autopsy

There are a number issues and challenges associated with VA: The VA tool (the classification system used, the questionnaire and the diagnostic technique employed), the process for data collection and the distribution of cause-specific mortality [72]. It is important that these issues are understood and considered with regard to the build of the prototype.

3.1.1 The Validity of Verbal Autopsies

As VA relies on the information provided by the caregiver to determine the cause for death with no clinical evidence to support, they may be subject to relatively high misclassification errors. "This can have a profound effect on the verbal autopsy estimate of the proportion of deaths due to a specific cause known as the cause-specific mortality fraction" [73,74]. *"Misclassification errors arise in two ways: (i) if a child who <u>did not</u> die from diarrhoea is classified as a diarrhoeal death or (ii) if a child who <u>did</u> die from diarrhoea is classified as a non-diarrhoeal death. These two issues outline the well known concepts of sensitivity and specificity. "Sensitivity being for the particular cause of death in this case diarrhoea, the proportion of the deceased whose cause of death was correctly identified as diarrhoea out of those who definitely died of diarrhoea. "The sensitivity being the proportion of death identified as not having diarrhoea among those who definitely did not die of diarrhoea"* [73]. Misclassification leads to either over or under estimation of the cause-specific mortality. In some studies misclassification has over estimated the CSMF by 5-12% [73]. However it is important to note that sensitivity and specificity, the standard evaluation metrics in epidemiology, are related to but not the same as the precision and recall metrics popular in NLP research.

The issue of misclassification has been widely debated [74-75] and there have been several attempts to find a solution[s] to address the issue. On examining the research conducted the broad conclusion drawn is the issue remains unsolved. On a positive note it was determined that specificity appeared to be more important than sensitivity in determining the accuracy of the VA tool. However the misclassification problem remains. There are two main reasons for this (i) there is a lack of validation studies. To explain, in a validation study, results from the verbal autopsy questionnaire are compared to the medical diagnosis known as the "gold standard", (ii) lack of information on the sensitivity and specificity within many of the VA tools. This could be explained by small sample data-set sizes used in many studies making sensitivity measures unreliable. So the learning from this must be that there needs to be greater effort placed on conducting quality validation studies and also developing information on the specificity and sensitivity within the tools.

Where quality validation studies have been carried out the results have proven to very valuable. Arguably one of the most practical methods is to conduct the study within a hospital setting where the VA questionnaire is completed with the care giver. A number of validation studies in hospital settings have been carried out [73,75-78]. These studies were undertaken using children in Bangladesh, Nicaragua and Uganda. The significance and value of these studies is that all three studies used the same unified standards. Often studies are completed with little regard to process, repeatability and comparability. These studies enabled sensitivity and specificity to be measured and the variation by country explained; highlighting the different disease patterns and also how the symptoms of the diseases were explained differently according to cultural traditions and local language [13]. All useful and valid findings which are directly relevant to establishing the accuracy and validity of VA.

However, validation studies in hospital settings do have limitations and these need to be understood and considered. The deceased may not have been representative of the general population and on death the care giver often learns the medical diagnosis and may be given the death certificate. This could affect the answers given at the VA interview. However, from a practical point of view, hospital validation studies are the only feasible method to validate a VA questionnaire [13]

This research became very relevant to the project as it clearly demonstrated the need to have validated results against the gold standard and also that the method of production of the gold standard was understood. The project was able to achieve this through use of building the first prototype using American medical discharge summaries where the gold standard was clearly cited and also through two sample of verbal autopsy data one provided by the London School of Hygiene and Tropical Medicine and the other by the Institute of Heath Metrics and Evaluation (IHME), University of Washington.

3.1.2 Standardisation of the Verbal Autopsy Questionnaire

In 2003, the WHO working with the Health Metrics Network (MHN) published a set of standards which outlined that different verbal autopsy questionnaires should be used based on age. There are three age groups under four weeks, four weeks to 14 years and 15 years and above [5]. Through research there is evidence that these standards have been adopted and are being used out in the field [25].

However despite concerted efforts led by the World Health Organization (WHO) to standardise the overall VA tools and coding procedures, due to the heterogeneity of both the process and its implementation this has yet to be achieved. [11].

To explain, there is no unified standard on the questionnaires used. They vary in both content and length, with some using open questions, some only closed questions and some a mixture of both [13]. Open ended questionnaires need to be coded by trained personnel and this incurs cost and time. However, the open format does enable a full account of the illness to be given which increases the probability of assigning an accurate cause of death. They are by nature tailored, so not to ask the care giver irrelevant questions or add further distress. Closed questions are more objective and often used with pre-defined algorithms. However, they have a number of disadvantages; inflexible as useful and relevant information may be omitted which aids the determination of cause of death and also the format could be viewed as lacking in sensitivity if not handled appropriately. This issue will only be addressed and resolved when standardisation in format and field operations are deployed and consistently used within countries and communities [5-6]. For the purposes of the project arguably the best approach to balance this issue is to ensure that the results are benchmarked against the gold standard.

3.1.3 Cultural Issues

Culture also affects the accuracy of the VA. The willingness of the care giver to agree to an interview, the description of the final illness and also the way that symptoms and disease is understood and described in the community are all important major contributing factors to the attainment of cause of death. Another factor is the attitude in the community towards particular causes of death. In some cultures some causes of death e.g. HIV may be under reported due to the stigma associated with this disease. Indeed, a very difficult issue to overcome. In relation to this project this presents real challenges. The learning being to ensure that the prototype from a NLP perspective has an ability to ensure that the all relevant information is extracted and included to support the cause of death diagnosis.

3.1.4 Data within Verbal Autopsy Questionnaire

When conducting the VA it is assumed that each cause of death has a set of observable features that can be recalled during the interview. Unsurprisingly VA performs best when it has distinct features that are not prevalent in other causes of death. If the information provided only gives a vague summary of symptoms and signs this can led to overlap and misclassification of cause of death. This affects all the interpretation methods; physician review, expert and data driven algorithms although arguably to greater and lesser extent (see Chapter 3: 3.1.7). Overall, studies have shown the VA has worked well for diseases such as measles, whooping cough, tetanus, cholera and dysentery as well as injury and cases of violence. Although they are less effective where symptoms are less specific e.g. HIV/Aids in children, malaria in adults and cancers [16,73,74,13].

3.1.5 Recording and Coding of Mortality Data

The agreed standard for recording mortality is worldwide through the use of the International Statistical Classification of Diseases and Related Health Problems which is now in its 10th Revision (ICD-10). ICD is the most widely used statistical classification system enabling the recording of diseases and signs, symptoms, abnormal findings, complaints, social circumstances and external causes of injury or diseases and is produced by the WHO [24]. The WHO stipulates the use of ICD in its most current revision for mortality reporting by its all Member States, currently 193 in total as of 2010 [25].

However, mortality reporting and coding is not without issue. Important research conducted in 2003 by Mathers et al on death registration produced some disturbing statistics. Death registration was available from 115 countries although in reality it was only complete for 64. Coverage of death registration varies enormously from nearly 100% in European Region but less than 10% in African Region. Some countries do not even use it: 75 member states including more than 90% of African countries have no information on cause of death available for any year after 1990 [4]. *"Health care prioritisation is conducted on the basis of perception, survey based information, levels of child mortality that are used together with model life tables, cause of death model and partial information from surveillance systems for some specific cause of death"* [4].

ICD-10 contains twice as many codes as ICD-9. Although in one perspective the revision was another step forward in improving mortality reporting providing access to over 14,000 codes and aids the tracking of new diagnoses, two main issues have developed as a consequence. One of comparability and also an increase in the use of coding categories for unknown and ill defined causes. The net result being that where data is available it has been harder to make comparisons over time on both a world/region and country basis and coding issues are still very much alive. Interestingly coding issues are not just a developing country problem. Although the problem of use defined codes exceeds 30% in countries such as Thailand and Sri Lanka, in some developed countries 10% of deaths are assigned ill defined codes [4].

There are a number of ways that this issue can be addressed, although none are a quick fix. Education of physicians and other key personnel involved the VA process on the importance of accurate and complete reporting on death certificates and avoidance of the use of ill defined codes is crucial. On a wider scale through public health policy making, further research is required to improve analysis of cause of death data. [5] Arguably it's the WHO that needs to play a pivotal role in facilitating and driving this forward.

3.1.6 Single vs. Multiple Cause of Death

Many VA studies assign a single cause of death, usually the underlying cause of death [13,79]. This means that the total number of causes of death is equal to the total number of deaths. On the surface this seems both sensible and intuitive. However, it is common that death is the result of more than one cause. *"For example a death primarily due to diarrhoea with concurrent pneumonia is indistinguishable from a death primarily due to pneumonia with concurrent diarrhoea. Therefore it is important when interpreting the results of a VA to understand whether multiple causes of death are allowed for in the coding"* [13]. This was a major consideration for the project; an assessment of what could be achieved either single or multiple cause of death based on the information obtained. This also affected the classification methods that could be used. This view is supported by the research of Reeves and Quigley [17].

3.1.7 Diagnostic Criteria

There is much debate over the accuracy and effectiveness of the diagnostic criteria [79]. To elaborate, considerable work on VA methodology has concentrated on emulating individual physician death certification, often glossing over the considerable variability and imprecision with which death certificates, the supposed "gold standard," are sometimes completed [3]. There has been debate on how to define a method as having high diagnostic accuracy. Research has shown that for use at *"the individual level high diagnostic accuracy exists if the sensitivity and specificity are at least 90%. At population level it occurs if the sensitivity is at least 50%, specificity at 90% and the CSMF within ±20% of the true value"* [73].

Physician review, expert and data driven algorithms have all been subject to validation studies and evaluation. The research is inconclusive in terms of gaining agreement on the best diagnostic methodology.

PCVA, given that it is conducted by a physician, appears to have validity and credence and it cannot be ignored that this is the most used method when conducting VA. Similar to medical history taking, physicians are local, aware of local customs/culture and also the disease patterns and symptoms within the area. However, research has shown these perceived benefits may cause PCVA not be the best method of establishing cause of death [72-74]. Issues have been raised over subjectivity, repeatability and the influence of bias. Also very importantly the time and costs implications incurred within this method limit its scalability.

Expert algorithms by their very nature provide a consensus of opinion from physicians. *"The algorithm is based on the symptoms deemed by the physicians to be essential, confirmatory or supportive in diagnosing cause of death"* [13]. Arguably, this method assists in dealing with the issue of inconsistency and addresses the time and cost issue. However, there are still concerns around the validity of this methodology. One of the main concerns being the inclusion of signs which are deemed as essential but have an inability to play an indiscriminating role within the process. To explain *"a VA study conducted in Kenya included fever in the expert algorithm for malaria but it had poor discriminating power as 93% of all malaria deaths and 86% of non-malaria* *deaths had fever"* [80]. Another issue has been the ability to include all the symptoms and scenarios which may led to cause of death so again, similar to physician review, this method again lacks scalability.

Computer techniques using data driven algorithms have also been used in the VA progress. There are a wide range of tools including logistics regression, neural networks and Bayesian probabilistic approaches. The results of several studies [12,81,82] have shown that data driven methods can perform as well as PCVA or expert algorithm, although there is an equal amount of research from the expert domain that states the contrary [80-81].

Data driven algorithms have been proven to be effective at deriving cause of death where the symptoms are specific although less effective when the symptoms are non specific such as pneumonia and malaria. In the main these algorithms do not use the information provided from the open question aspect of the questionnaire [6,80,83]. Although excluding this information may make the data less subjective the disadvantage is that important information may be missed [84]. The general consensus is that the information contained in the "open" sections is considered to be of greater value and importance than the information within the closed [6,16]. The lack of standardised VA questionnaires limits the ability to build and standardise the algorithms and also there is these is also an argument that to increase their effectiveness they benefit from being given a context-specific approach [13]

More recently there has been development in probabilistic approaches. The research of Byass using Bayesian approaches has provided some promising work through the development of the "InterVA model" [82,85-87]. The approach has the ability to establish individual cause of death by using the symptom level data recorded in the VA. The method calculates the likelihood of each cause and displays up to three possible causes of death. The ability to assign multiple causes of death, having the ability to take into account local disease prevalence and through its application appears to perform well against physician review makes this work attractive. However, it is not without criticism. *"The method is considered of limited use at the individual level and the lack of a gold standard with which to validate diagnoses has restricted its application"* [6].

Another probabilistic approach was developed by King and Lu [88-89] which directly estimates CSMF without individual case of death attribution, Data on the symptoms provided by the care giver along with the cause of death are obtained from health facilities and the cause of death distribution is estimated in the population from the symptom data available. The method has more complexity than InterVA and research conducted in China and Tanzania has shown that it performs well on ascertain probability levels. However, there is one major drawback in that it depends on the availability of high quality health facility based mortality data. Herein lies the problem; there just isn't enough of it. However, this work takes an interesting new approach and has encouraged further research in this area.

Murray et al have combined the works of King and Lu and Byass with the InterVA method to develop the "symptom pattern method" [79,12]. To work, a dataset where the true cause of death is known is needed so that specific symptoms given a specific cause of death can be established and quantified. From this population and individual levels cause patterns to be determined from the second data set from the population profile. The method was validated using a sample of 2000 deaths in China where the gold standard was available [79]. The results showed that this method outperformed PCVA at both population level and individual level. This is again promising work but similar to work of the King and Lu to be effective it requires a substantial dataset of symptom level data and a high standard of facility based data. Therefore, to enable more research to be undertaken on this and similar methodology, more volume and quality data is required.

3.1.8 Conclusion: Looking to the Future

It is evident that the VA process, the recording and reporting of is a very complex and complicated task.

One of the most challenging aspects of developing VA is the breadth of purpose that the information is used. From establishing individual causes of death, population cause of death, infectious disease outbreaks and to assist with global and national cause specific mortality estimates. This has had major effect on consistency, compatibility and adequacy of the VA tools and their development. The net result being that despite the copious studies and literature, different research favours particular methodologies. Thus, it is fair to make two statements: firstly, the literature provides an inconsistent picture and secondly, based on this it is unlikely that in the near future a one-fit-all methodology will emerge.

However, what cannot be disputed is to that to move VA methodology forward certain issues need to be resolved. Standardisation of the documentation and field operation procedures are key as well as education of and improving coding standards. Sample data sets to evaluate methodologies need to be larger and also have the associated gold standards. To produce quality research both quality and volume are vital. Only then will further automation and computational approaches move forward.

Despite the known issues the overriding consensus within the medical and academic world is that VA still is the most appropriate and useful method for documenting cause of death where there is no medical supervision.

3.2 Data and System Preparation: Selection of the Data

Some key decisions needed to be made on the selection of data to move forward to modelling and prototype phase.

3.2.1 Discharge Summary Sample

On preparing the discharge summaries ready to process into GATE a number of key discoveries were made. Firstly, the sample from The University of Pittsburgh Medical Centre included 81 “progress reports” within the overall 180. Progress reports detail the ongoing treatment of a patient whilst at the hospital. On examining the progress reports a decision was made to exclude them from the final data sample going forward to build the prototype. This was for a number of reasons; in a high percentage of the reports it was difficult to ascertain the reason for the patient’s admittance and actual diagnosis, therefore the gold standard was ambiguous. This was a major concern as all research conducted and documented in Chapter 1 had pointed to the need to have the associated gold standards to support a final diagnosis. The concern was that if a lay person interpreted the diagnosis it could inject bias or incorrect information into the results, so for the avoidance of any doubt they were extracted from the sample. The sample then stood at 269. Secondly, on further scrutiny, when the sample loaded into GATE another issue became clear; the diversity of the illnesses and diseases within the discharge summaries. Although there was no wish or desire to “tamper” with the 269 sample any more, it became evident that the sample contained many single diseases/illnesses/complaints/procedure occurrences. The scope was extremely broad, examples being from requests for sterilization, shortness of breath, various types of cancers, circuit video electroencephalographic monitoring, carbon dioxide poisoning to name just a few. To enable a classifier to be built successfully there needs to be more than a single case to “train” the data. As a result what was a sample of 269 became a reduced sample of 16. Within this sixteen, three classes were obtained. A sample of 8 patients who had Pneumonia, 3 who had Chronic Obstructive Pulmonary Disease (COPD) and 5 with Coronary Artery Disease (CAD). Although disappointing from a classification perspective, the data sample still enabled the full prototype to be built and tested.

3.2.2 Ghana Verbal Autopsy Sample

This data set provided two opportunities. The word document clearly showed a structured format, including both open and closed questions. Where there were open questions, known within the document as “the story of illness” there was opportunity to process this information in GATE. By doing this the prototype process would mirror that of the discharge summaries. The story of illness section of the questionnaire is where the interviewer invites the mother to give her personal account of the pregnancy, the birth and where appropriate (if not a still birth) the events leading up to the baby's death including any signs, symptoms or treatments that took place. The other opportunity was to use the csv file, removing any non relevant attributes and then upload straight into WEKA; in this respect the annotation phase via GATE would be removed. This would recreate current practice where it has been acknowledged that the majority of data driven algorithms discount the information in the open sections of the questionnaire. A decision was made to do both exercises and compare the results. Unfortunately, with such a small sample, it would be unlikely that any significant findings could be derived, although it would illustrate the process. To really benefit a larger test set would be required to be tested on the classifier. The gold standard cause of death diagnoses were provided for this sample and it was found to have two deaths from severe infection, one premature, one congenital abnormality and the other was unexplained see Appendix L.

3.2.3 IHME Verbal Autopsy Sample

The complete sample of 1592 verbal autopsies would be used, once the final class attribute was converted from number (integer) to an identifier (e.g. 1 to x1) no errors were picked up during initial data load and no missing values were found. Through communication with the "gatekeepers" of this data it was established that some of the designated symptoms were actually non symptom attributes which needed to be removed during the cleaning phase to enable optimum processing results.

3.3 Cleaning the Data

All the data sets required some aspect of data cleanse before processing, to a greater or lesser extent.

3.3.1 Discharge Summary Sample

The 81 progress reports were removed from the data set. On initial load into GATE it was found that the discharge summaries failed to annotate effectively. Within the GATE, documentation was supposed to be able to be case agnostic, however it was found that the case sensitivity was not working effectively so to combat this all the discharge summaries were changed into lower case. This was achieved by writing a python program see Appendix P.

3.3.2 Ghana Verbal Autopsy Sample

For each of the five VA's the story of the illness section was manually extracted from each document and built into a raw text file for processing into GATE. The CSV file was checked and no errors or missing values were found. In total there were 246 attributes within the data set, and on checking the data set 12 attributes were removed before processing, leaving a total of 234. The attributes removed were all the unique identifiers such as woman id and infant id, batch number, interviewer number. If these had remained in the csv file then the classifier would have predicted on these unique attributes and therefore the results would have been incorrect. Within the csv file there were some special values to understand "9" and "999" meaning not applicable, "8" and "888" both meaning not known and "0" meaning none.

3.3.3 IHME Verbal Autopsy Sample

The CSV file contained 142 "symptoms" (although if using the correct terminology they should be referred to as "attributes") in total. After gaining some additional information on the attributes within sample, 10 attributes were removed from the data set. These 10 attributes were deemed as noise and best removed from the data set to ensure the most accurate results. Symptom 2 was removed as it was an age variable, symptoms 27, 40, 45, 73, 77, 81, 83, 90, and 138 all describe the duration of symptoms listed elsewhere in the questionnaire and symptom 140 was a location variable. Another aspect to the data was to identify the special values within the csv file; "99" meant "did not know" and "-1" meant no response. This was important to understand when reviewing the results from the classifier.

3.3.4 SNOMED-CT Data File

The SNOMED-CT file needed to undergo some basic but very crucial cleaning. As discussed previously on receipt of the files it was very evident that each concept within the raw data file was annotated with a hierarchy description. These were removed from the raw text file by the python program. After this was completed the file then needed to be changed into lower case. Fig 3.1 a snapshot of the SNOMED-CT file once the hierarchy descriptions have been removed and clearly shows that the file has a mix of both upper and low case word structure. Without correction this would have caused annotation issues when the file was built into a gazetteer to be processed within GATE.

Once the data had been fully examined, assessed and then necessary cleaning had been completed, satisfaction was reached that the data was in a quality format suitable to be put forward to be loaded in both GATE and also WEKA. So the final data sets were as follows, see Table 3.1. In total there were 16 US discharge summaries. These went through each stage of the completed prototype GATE, Python and WEKA. The Ghana verbal autopsies (story of the illness section) again through all stages of the prototype. The 1592 IHME verbal autopsies and the Ghana verbal autopsies (same 5) but in format 2, i.e. the CSV file went through the WEKA process only.

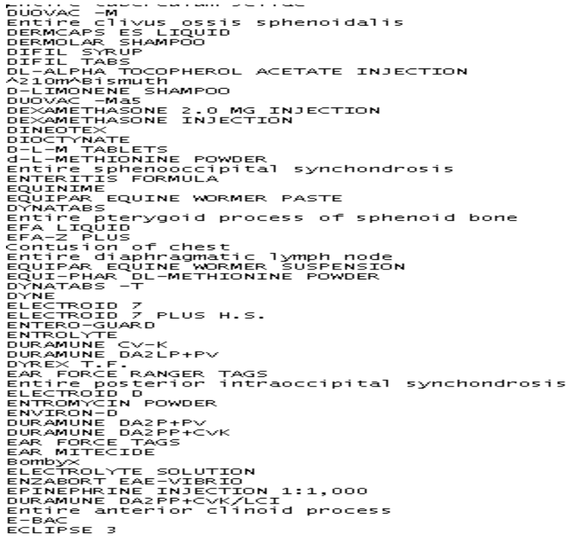

```
DUOVAC -M
Entire clivus ossis sphenoidalis
DERMCAPS ES LIQUID
DERMOLAR SHAMPOO
DIFIL SYRUP
DIFIL TABS
DL-ALPHA TOCOPHEROL ACETATE INJECTION
^210m^Bismuth
D-LIMONENE SHAMPOO
DUOVAC -Ma5
DEXAMETHASONE 2.0 MG INJECTION
DEXAMETHASONE INJECTION
DINEOTEX
DIOCTYNATE
D-L-M TABLETS
d-L-METHIONINE POWDER
Entire sphenooccipital synchondrosis
ENTERITIS FORMULA
EQUINIME
EQUIPAR EQUINE WORMER PASTE
DYNATABS
Entire pterygoid process of sphenoid bone
EFA LIQUID
EFA-Z PLUS
Contusion of chest
Entire diaphragmatic lymph node
EQUIPAR EQUINE WORMER SUSPENSION
EQUI-PHAR DL-METHIONINE POWDER
DYNATABS -T
DYNE
ELECTROID 7
ELECTROID 7 PLUS H.S.
ENTERO-GUARD
ENTROLYTE
DURAMUNE CV-K
DURAMUNE DA2LP+PV
DYREX T.F.
EAR FORCE RANGER TAGS
Entire posterior intraoccipital synchondrosis
ELECTROID D
ENTROMYCIN POWDER
ENVIRON-D
DURAMUNE DA2P+PV
DURAMUNE DA2PP+CVK
EAR FORCE TAGS
EAR MITECIDE
Bombyx
ELECTROLYTE SOLUTION
ENZABORT EAE-VIBRIO
EPINEPHRINE INJECTION 1:1,000
DURAMUNE DA2PP+CVK/LCI
Entire anterior clinoid process
E-BAC
ECLIPSE 3
```

Fig: 3.1: A snapshot of SNOMED-CT Concept File

DATA SET	SIZE	FILE FORMAT	GATE	PYTHON	WEKA
Discharge Summaries	16	Plain Text	✓	✓	✓
Ghana Verbal Autopsies	5	Plain Text	✓	✓	✓
Ghana Verbal Autopsies	5	CSV File	✗	✗	✓
IHME Verbal Autopsies	1592	CSV File	✗	✗	✓

Table: 3.1: Final Data Selection, Preparation and System Usage

3.4 System Preparations

Before moving onto the modelling stage of the project it is important to advise the preparations that were undertaken from a systems point of view to move forward with the prototype.

3.4.1 GATE

GATE was not a system that had been part of the syllabus of the course therefore a practical understanding of the functionality, the “behaviour” and abilities of GATE needed to be acquired before a prototype could be built. The knowledge and understanding came from the various learning tutorials on the web and a short one hour workshop which took place at the University on the key but basic features of the system. The overall experience with regard to "setting up" GATE with a view to presenting it with text "to engineer" was quite a painful one, exacerbated further when one is not familiar with NLP terminology and practices. It is fair to say that assumptions are made that the user already has a level of NLP understanding to set up the system ready for use. There was much learning from errors made and these are expanded upon in the subsequent chapters.

3.4.2 Python

It was intended that the GATE tool would be used to identify, annotate and extract the medical concepts from each of the medical text documents. At data preparation stage it was clear that GATE had some limitations in that it was unable to output the results from the annotation phase of the prototype. As a result an "add in" process needed to be built for the prototype to work accordingly. A python program was written which read the contents of all the medical text documents one by one and output the results into ARFF format. With this format produced it was then loaded straight into WEKA. If the program had not been built then it would have been a manual extraction process which would have been unscalable on a greater volume of verbal autopsy documents. Python was also used to clean the concept files.

3.4.3 WEKA

In terms of system preparation, the python program closed the gap in the process for the prototype build where plain text files were the source of data producing an accurate ARFF file for upload to WEKA. Decisions were made on what algorithms to use. Within the course only decision trees had been taught so there was a high degree of background reading required to understand the available array of algorithms, their functionality and purpose to enable an informed decision to be made on which ones would be most appropriate. This was arrived at by considering the current practices in VA interpretation and reading papers on machine learning against the backdrop of the data set and purpose of the project.

3.5 Modelling: Prototype Model

In its basic form the prototype had six key steps: Firstly, the acquisition of the medical text documentation and medical terminology. Then assess the format of each of the data sets. The next consideration was the pre-processing of the data to enable the successful load into GATE, followed by a Python program to extract the concept terms. Finally to then load the data (concepts now taken the form of attributes) into WEKA and build various classifiers to establish some results, see Fig 3.1. There were some changes to the model based on data format which were alluded to in the previous chapter and will be discussed further in this chapter. From building and using this prototype an evaluation could then be undertaken into the verbal autopsy process to understand the issues and challenges from a computational perspective.

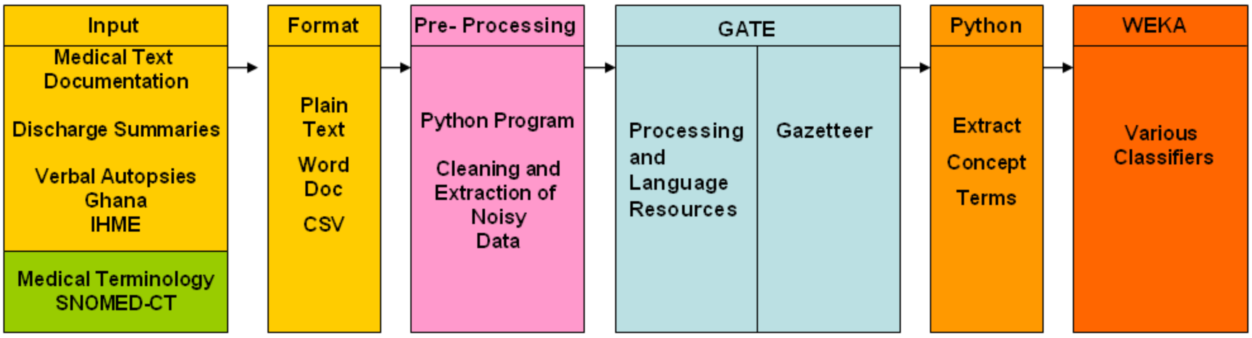


Fig 3.1: Basic Prototype Model

3.5.1 Classifier/Algorithm Selection

Three rule based classifiers were used to baseline the results; ZeroR, One R and J-Rip. OneR is as it states a simple 1 parameter classifier. ZeroR predicts the majority class if nominal or the average value if numeric; in the case of this project it predicts the major class. Finally J-Rip implements RIPPER which is an acronym for repeated incremental pruning to produce error results [90].

After the baseline was obtained a further set of classifiers were used. Through the literature research it was established which methods have been used previously. The author wanted to use a breadth of learning algorithms types so from WEKA the following were chosen: Naïve Bayes is a standard probabilistic classifier which has proven a popular approach in verbal autopsy, J48 a decision tree not a common methodology but an interesting choice, MultilayerPerceptron a neural network which works on back propagation, LogisticR a regression method and finally Adaboost.M1 which is a method that combines multiple models and weightings.

Usually in a data mining project part of the sample would be used as the training set and then the remainder for testing or a new set of data would be used. Unfortunately due to the small data sets this was not possible. Although not ideal, to mitigate all the classifiers were built where possible using the cross validation function.

3.5.2 Initial Steps

With the data understanding and preparation stages completed successfully the next step was to start to "program" GATE to be able to carry out the tasks correctly and accurately. This involved a three stage process;

1. Building an annotation pipeline in GATE
2. Construction and loading of the SNOMED-CT file to build a "gazetteer" in GATE.
3. Building a set of corpora to load into GATE for annotation

The annotation pipeline was built using "ANNIE" within GATE. Although ANNIE consists of a wide range of processing resources the requirements for this project were a tokenizer, sentence splitter and also the ability to build a gazetteer. The concept of using the gazetteer was that once functioning when run over each corpus, it would annotate the text tokens when a match was found.

Although this seemed a straight forward process when building the annotation pipeline there were many options for different processing resources tools available. The support documentation was comprehensive but lacked intuitiveness when read by a complete novice wishing to undertake such a task. It was not clear which processing resources would be best served or the order to load in the processing resources for maximum benefit. This issue was resolved through trial and error.

The Gazetteer build also proved a challenge. Although the support documentation explained in detail about what a gazetteer was and also there were pre-formatted gazetteers already contained within ANNIE there were scant instructions of how to set up a new gazetteer. To resolve the pre-formatted gazetteer locations were identified and the new gazetteer containing all the SNOMED-CT concepts was populated in the same location to enable the file to be read.

To enable the use of Language Processing within GATE then a number of corpora needed to be built for each data set. The following corpora were produced:

Corpus 1: US discharge summaries. Size: 16 (omitting all but 3 disease findings)

Corpus 2: USA discharge summaries Pneumonia: Size: 8

Corpus 3: US discharge summaries Chronic Obstructive Pulmonary Disease. Size: 3

Corpus 4: US discharge summaries Coronary Artery Disease. Size: 5

Corpus 5: Ghana verbal autopsies. Size: 5

Each of these was saved into a separate datastore within GATE to enable the fast retrieval of each corpus when required. This was fortunately a relatively straightforward process.

3.5.3 Initial Data Load into GATE

On initial data load into GATE with Corpus 1 a number a key issue arose. When the corpus was run over the gazetteer the concept annotations were extremely small in number, less than 3% of the corpus, see Table 3.2. Through investigation it was established that case sensitivity was the issue. Despite GATE being documented as being case agnostic clearly there were some issues. As a result the gazetteer and all the discharge summaries were changed into lower case so that the matching between the corpus and the gazetteer would be optimized. Once completed and with satisfaction that the compatibility issue had been resolved the first prototype build could move forward.

Discharge No:	Tokens	Annotations	Percentage
22	5945	104	1.75%
23	3180	78	2.45%
28	3660	82	2.24%
36	5460	104	1.90%
51	5360	91	1.70%
74	5725	112	1.96%
78	7080	131	1.85%
88	4640	78	1.68%
39	960	50	5.21%
95	2284	105	4.60%
67751445	2570	88	3.42%
20	572	53	9.27%
34	317	33	10.41%
50	574	39	6.79%
96	1208	112	9.27%
119	1557	155	9.96%
Total	**51092**	**1415**	**2.77%**

Discharge No:	Tokens	Annotations	Percentage
22	5945	1062	17.86%
23	3180	609	19.15%
28	3660	656	17.92%
36	5460	941	17.23%
51	5360	890	16.60%
74	5725	926	16.17%
78	7080	1210	17.09%
88	4640	751	16.19%
39	960	268	27.92%
95	2284	742	32.49%
67751445	2570	658	25.60%
20	572	171	29.90%
34	317	82	25.87%
50	574	149	25.96%
96	1208	374	30.96%
119	1557	492	31.60%
Total	**51092**	**9981**	**19.54%**

Table 3.2: Concept annotations shown in GATE. Table on left shows results where no changes make to case. Table on right shows results after case sensitivity has been removed

3.3 Prototype Build

Due to the tardiness of the verbal autopsy samples the discharge summaries where used first. The benefit of the discharge summaries was seen as both their size and also the format which would allow the data to go through every stage of the prototype process enabling a complete evaluation to be obtained and documented.

3.3.1 Discharge Summary Prototype

In total three prototypes were built and evaluated. To refresh the memory, this data set consisted of 16 discharge summaries. Within the set there were 3 classes, 8 cases of where a patient had been diagnosed with Pneumonia, 5 cases of Coronary Artery Disease and 3 cases of Chronic Obstructive Pulmonary Disease.

Prototype 1:

The concept behind the very first prototype was to process sample very much in the same way as a physician coded autopsy would be undertaken. Although it is acknowledged that author has no medical background. In this prototype the GATE process was removed and in place a manual human process was inserted. The text within all 16 documents was read and then the key medical signs and symptoms within each document were manual highlighted and a manual count of the frequency of these words was documented this was completed to form the basis of the ARFF file for the classifier purposes. In total 21 symptoms of disease (attributes) which were extracted from the 16 discharge summaries see Fig 3.3 based on frequency of the words used. From this an ARFF file was built using notepad in preparation for the load into WEKA for the prototype process see Appendix P for the actual ARFF file produced. To see the complete process refer to Fig 3.4

Cough, coughing, pleural, effusion, lobe, sputum, fluid, WBC, lung, chest, angina, shortness of breath, hypertension, infarction, blood, pressure, artery, catheterization, chest x-ray, fever, chills

Fig 3.3 Prototype 1: The 21 identified signs and symptoms

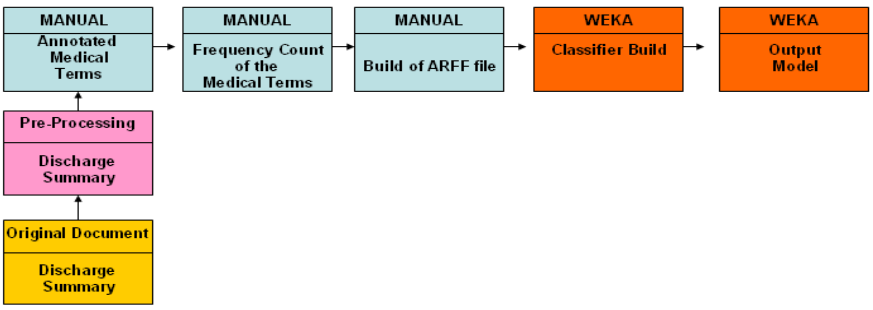


Fig 3.4 Prototype 1: Process Model for the Discharge Summaries

Prototype 2:

In this prototype the full automated process was applied. The corpus of the 16 discharge summaries was loaded into GATE. The original 21 medical signs and symptoms from Prototype 1 were observed in GATE to see if they received a mark-up in GATE i.e. to ascertain if medical terms chosen in the first prototype were in fact recognized SNOMED-CT concepts; there was a match with the terms in both the discharge summary and the gazetteer (see Appendix Q for an example of annotated summary).Of the 21 original, 7 were removed as they were not recognized SNOMED-CT concepts. These were lobe, lung, infarction, chest, hypertension, pressure, and artery. It was found that SNOMED-CT does not recognise single word plurals so "chills" was changed to "chill" a recognized concept. Where the gazetteer identified the FULLY SPECIFIED NAME present the original term was replaced. For example "infarction" became "myocardial infarction". The new list of concepts can be seen below in Fig 3.5.

Cough, coughing, pleural, effusion, lobe, sputum, fluid, WBC, angina, shortness of breath, blood pressure, catheterization, chest x-ray, fever, chill, pulmonary hypertension, myocardial infarction, pleural effusion, pericardial effusion, green sputum, cardiac catheterization, renal stenosis, coronary artery and white sputum.

Fig 3.5 Prototype 2: The 24 concepts

This increased the overall SNOMED-CT annotations to 24. The Python program was then run to extract the annotated SNOMED-CT concepts and to count the frequency that they occurred within the text. For details of the python code see Appendix R. The program produced the output file in ARFF format and the annotations then became the attributes for the classifier (see fig 3.6).

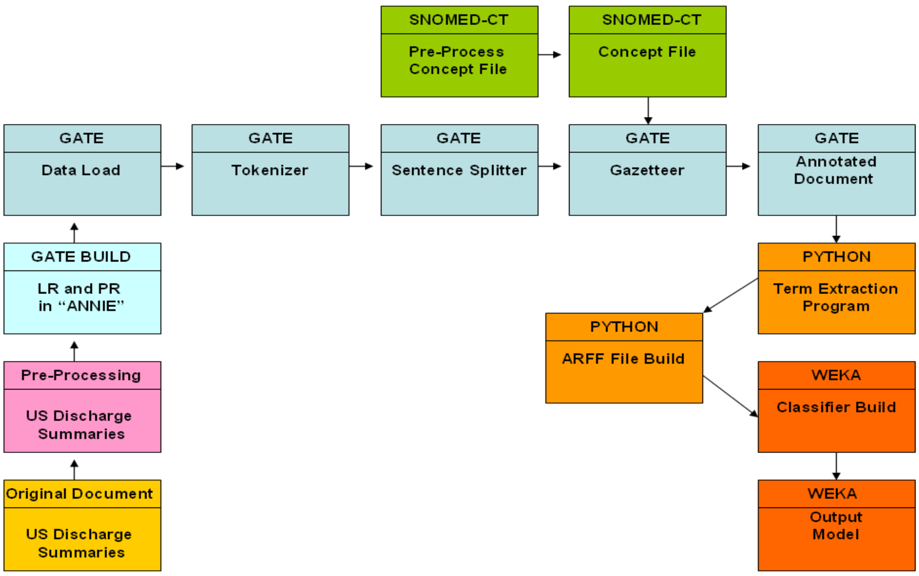


Fig 3.6 Prototype 2: Process Model for the Discharge Summaries

Prototype 3:

This prototype again used the full automated process as shown in Fig 3.6. The difference with this prototype is that the gazetteer was run over all the documents and every SNOMED-CT concept that was annotated in GATE using the SNOMED-CT gazetteer was extracted. This produced a significantly larger number of SNOMED-CT concept annotations in total 9981, see Table 3.3.

Discharge No:	Tokens	Annotations	Percentage
22	5945	1062	17.86%
23	3180	609	19.15%
28	3660	656	17.92%
36	5460	941	17.23%
51	5360	890	16.60%
74	5725	926	16.17%
78	7080	1210	17.09%
88	4640	751	16.19%
39	960	268	27.92%
95	2284	742	32.49%
67751445	2570	658	25.60%
20	572	171	29.90%
34	317	82	25.87%
50	574	149	25.96%
96	1208	374	30.96%
119	1557	492	31.60%
Total	**51092**	**9981**	**19.54%**

Table 3.3 Prototype 3: Discharge Summaries: SNOMED-CT Annotation Results

3.3.2 Ghana Verbal Autopsy Prototype

Prototype 1: Story of Illness

This prototype again used the full automated process as shown in Fig 3.7. When loaded into Gate the corpus which contained the free text section within the document, the story of the illness a total of 2658 tokens and within this 551 SNOMED-CT concepts were achieved see Table 3.4.

VA Number:	Tokens	SNOMED-CT Concept	Percentage
VA SOI 1	427	125	29.27%
VA SOI 2	383	92	24.02%
VA SOI 3	244	52	21.31%
VA SOI 4	635	116	18.27%
VA SOI 5	969	166	17.13%
Total	**2658**	**551**	**20.73%**

Table 3.4: Prototype 1: Ghana Verbal Autopsies: SNOMED-CT Annotation Results

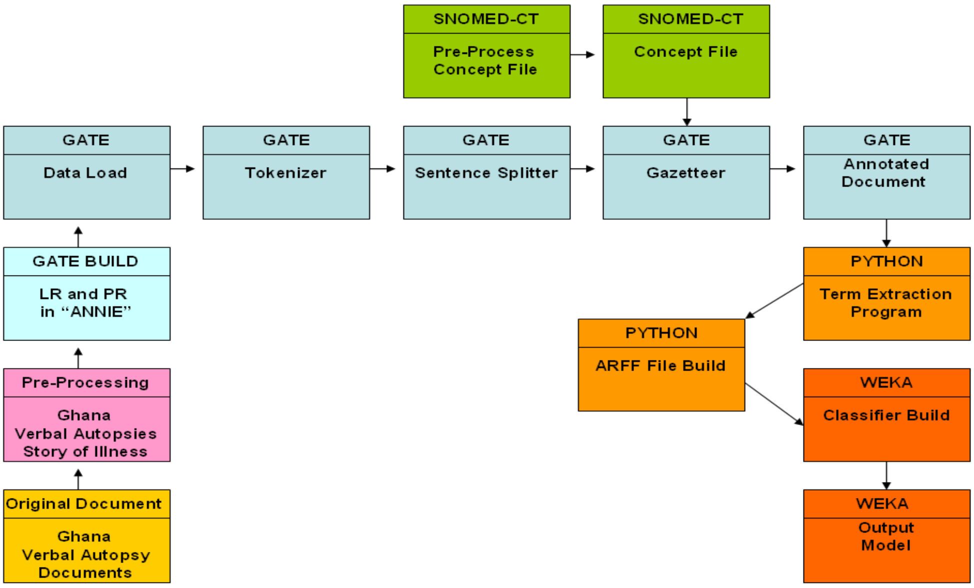


Fig 3.7 Prototype 1: Process Model for the Ghana Verbal Autopsy Prototype Format 1

Prototype 2: CSV format

The CSV file detailed the responses in the full questionnaire. In total there were 234 attributes loaded into WEKA to run the classifiers. The process can be seen below in Fig: 3.8

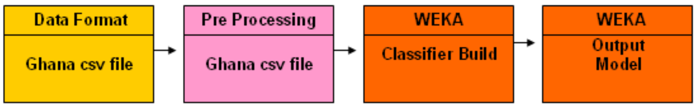


Fig 3.8 Prototype 2: Process Model for the Ghana Verbal Autopsies

3.3.3 IHME Verbal Autopsy Prototype

The CSV file detailed the responses in the full questionnaire. In total there were 132 attributes loaded into WEKA to run the classifiers. The process can be seen below in Fig 3.9.

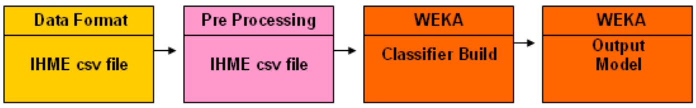


Fig 3.9: Process Model for the IHME Verbal Autopsies

In the next chapter the results are provided together with a complete evaluation of the project assessed against the aims and requirements outlined at project commencement.

Chapter 4: Evaluation

4.1 Introduction

When this project was embarked upon its intentions was to look at two broad areas;

(i) To research the verbal autopsy process to a gain a real insight and examination of this process understand the issues and challenges both of manual effort and computational methods.

And,

(ii) To illustrate and document these through the build and delivery of a prototype which sought to replace the role of both “coder” and “physician” to establish an accurate cause of death.

Before discussing the results of each prototype in detail, it is pertinent to provide some general evaluation against the minimum and additional requirements as set out in 2.5.

The research aspect of this project fulfilled the first three aims of the minimum requirements which proved to be an extremely challenging process. The primary reason why verbal autopsy is in place is that the countries that use it are without the infrastructure and financial resources to support them to build vital registration systems which in the western world are taken for granted. Although verbal autopsy is a seen as the best method to address this shortfall the whole process it is fraught with issues which makes it a very complicated problem space to examine and document. Overall it’s a fragmented problem space and despite efforts going back over 30 years and significant organisations involved such as the WHO there has been little traction in gaining consistency within the process. As a result countries and indeed regions in countries all conduct the process differently and as such any research findings reflect this. It has been reported that in some DSS sites questionnaires have not changed in over 10 years due to the expense of updating and retraining [93]. In short, progress is slow and painful. Although significant research papers were found on verbal autopsy, very few examined how to move this issue forward from a computational perspective; where evidence for this was found it was documented in 3.1.7. What really resonated throughout the research stage was the lack of agreement on which computational approaches are best or should be further explored. Although a personal view from the research it seems that Physician Review and Expert Algorithm (again by Physician) are both seen as being more clinically credible than computational methods, even though they are not without fault. In fairness to this statement what isn’t being implied is that the health profession is not interested in new methods but the constraints around sourcing quality data and in the volume required mean that none of the computational approaches have been tested robustly enough to warrant widespread clinical credibility.

With regard to researching the terminological systems, this was a minefield of ambiguity when trying to establish the differences between each system, their features and characteristics. Research papers often cited them as being used but again there were very few that explained the rationale of why they were being used. To enable a comprehensive covering of the subject the research included moving into the area of medical informatics. What was also was discovered was that although ICD-10 should be used as the core terminology reference in the field, cut downs of the terminology were used and in some areas not used at all, with preference to other terminologies or practices [91].

When examining the approaches of the extraction or recognition of natural language within the medical domain what was very clear is the volume of research that has been conducted on electronic patient records, including an array of research on extracting the free text narratives from discharge summaries. Through the literature search on verbal autopsies on medical text extraction it was a completely different picture. In fact the research advised that from a computational perspective the free text aspects of the verbal autopsy questionnaire were excluded from processing.

The background research conducted, although challenging to fuse together, provided an excellent foundation to build the prototype. The issues around gaining sample medical text are already well documented within the body of the report and also in the project reflections in Appendix A, so no further comment is required. The build of the prototype very much assisted in drawing out the issues associated with this process from a computational perspective. Although not without its challenge again the prototype was built and through its implementation medical text (both discharge summaries and verbal autopsies) were annotated, extracted and classified. As a result, conclusions are drawn and avenues for enhancement are advised. A more detailed evaluation of the results from each prototype and the systems/programs used to undertake this work are recorded in the remaining sections of this chapter.

4.2 Prototype Results

A benchmark needed to be applied to the results to determine its ability to predict accurately the cause of death. Although there is not definite agreement on this among experts, again another consistency issue within the overall process, the view of Anker which is supported by many experts; in order for a verbal autopsy classifier to be useful for classifying the death of an individual, it should be able to classify a death due to a disease with a sensitivity (true positive rate) near 90%; or in other words, it must have a generalization error (1-specificity) less than or equal to 10% [16].

"Sensitivity" and "specificity" are statistical measures of the performance. Sensitivity is often also known as the recall rate and measures the proportion of actual positives which are correctly identified as such; the percentage of people who are correctly identified as having a disease. Specificity measures the proportion of negatives which are correctly identified; the percentage of well people who are correctly identified as not having the disease [92,93].

$$\text{specificity} = \frac{\text{number of True Negatives}}{\text{number of True Negatives} + \text{number of False Positives}}$$

$$\text{sensitivity} = \frac{\text{number of True Positives}}{\text{number of True Positives} + \text{number of False Negatives}}$$

To explain in layman's terms; the "True Positive Rate" is the cases of disease where the classifier shows that they have the disease and they actually do. The "False Positive Rate" is the cases of disease where the classifier shows that they have the disease when actually they do not. The below table 4.1 explains the terms succinctly.

		Actual Disease	
		Disease Present	Disease Absent
Test	Positive	Disease Present + Positive result = True Positive	Disease absent + Positive result = False Positive
Result	Negative	Condition present + Negative result = False (invalid) Negative	Condition absent + Negative result = True (accurate) Negative

Table 4.1 Explaining Disease Result Outcomes: Source: http://encylopedia.the freedictionary.com/sensitivity

Other classifier measurements that will be examined are "Precision" which is the number of true positives correctly labeled as belonging to the class. The equation below makes this a simple concept to understand.

$$\text{Precision} = \frac{tp}{tp + fp}$$

"Recall" which is the total number of true positives divided by the total number of elements that actually belong to the positive class i.e. the sum of true positives and false negatives which were not labelled as belonging to the positive class but should have been. In this context Recall also refers to as the true positive rate. Therefore relating back to the above the true negative rate is also known as the "specificity" and false negative rate is known as the "sensitivity" [92,93].

$$\text{Recall} = \frac{tp}{tp + fn}$$

Before the results are discussed it is recognised that due to small sample size the validity of the results in terms of offering definite and exacting conclusions are problematic. A larger sample would have significantly increased the statistical validity of the findings. However, the results despite this provide an interesting proof-of-concept and again bring out the computational issues and challenges associated with the verbal autopsy process. The complete set of classifier results from WEKA can be found in Appendix S-U. Although a summarized version for each data set is below.

Discharge Summaries:			
Number of Attributes	21	24	146
Total Number of Instances	16	16	16
ZeroR Cross Validation	**Prototype 1**	**Prototype 2**	**Prototype 3**
% Correctly Classified Instances	50%	50%	50%
% Incorrectly Classified Instances	50%	50%	50%
OneR Cross Validation	**Prototype 1**	**Prototype 2**	**Prototype 3**
% Correctly Classified Instances	62.5%	75%	75%
% Incorrectly Classified Instances	37.5%	25%	25%
J-Rip Cross Validation	**Prototype 1**	**Prototype 2**	**Prototype 3**
% Correctly Classified Instances	43.75%	62.5%	68.75%
% Incorrectly Classified Instances	56.25%	37.5%	31.25%
J48 Cross Validation	**Prototype 1**	**Prototype 2**	**Prototype 3**
% Correctly Classified Instances	75%	75%	68.75%
% Incorrectly Classified Instances	25%	25%	31.25%
Naïve Bayes Cross Validation	**Prototype 1**	**Prototype 2**	**Prototype 3**
% Correctly Classified Instances	62.5%	68.75%	43.75%
% Incorrectly Classified Instances	37.5%	31.25%	56.25%
MultiLayerPerceptron Cross-Val	**Prototype 1**	**Prototype 2**	**Prototype 3**
% Correctly Classified Instances	50%	50%	50%
% Incorrectly Classified Instances	50%	50%	50%
AdaboostM1 Cross Validation	**Prototype 1**	**Prototype 2**	**Prototype 3**
% Correctly Classified Instances	75.0%	100%	100%
% Incorrectly Classified Instances	25.0%	0%	0%
Logistic R Cross Validation	**Prototype 1**	**Prototype 2**	**Prototype 3**
% Correctly Classified Instances	43.75%	50%	50%
% Incorrectly Classified Instances	56.25%	50%	50%

Table 4.2 WEKA Results: Discharge Summaries

Ghana:		
Number of Attributes	51	234
Total Number of Instances	5	5
ZeroR Cross Validation	**SOI**	**CSV**
% Correctly Classified Instances	40%	40%
% Incorrectly Classified Instances	60%	60%
OneR Cross Validation	**SOI**	**CSV**
% Correctly Classified Instances	40%	40%
% Incorrectly Classified Instances	60%	60%
J-Rip Cross Validation	**SOI**	**CSV**
% Correctly Classified Instances	40%	40%
% Incorrectly Classified Instances	60%	60%
J48 Cross Validation	**SOI**	**CSV**
% Correctly Classified Instances	60%	60%
% Incorrectly Classified Instances	40%	40%
Naïve Bayes Cross Validation	**SOI**	**CSV**
% Correctly Classified Instances	100%	100%
% Incorrectly Classified Instances	0%	0%
MultiLayerPerceptron Cross-Val	**SOI**	**CSV**
% Correctly Classified Instances	100%	100%
% Incorrectly Classified Instances	0%	0%
AdaboostM1 Cross Validation	**SOI**	**CSV**
% Correctly Classified Instances	60%	60%
% Incorrectly Classified Instances	40%	40%
Logistic R Cross Validation	**SOI**	**CSV**
% Correctly Classified Instances	100%	100%
% Incorrectly Classified Instances	0%	0%

Table 4.3 WEKA Results: Ghana VA's

IHME:	
Number of Attributes	132
Total Number of Instances	1592
ZeroR Cross Validation	
% Correctly Classified Instances	11.6834%
% Incorrectly Classified Instances	88.3166%
OneR Cross Validation	
% Correctly Classified Instances	19.0327%
% Incorrectly Classified Instances	80.9673%
J-Rip Cross Validation	
% Correctly Classified Instances	27.6382%
% Incorrectly Classified Instances	72.3618%
J48 Cross Validation	
% Correctly Classified Instances	26.8844%
% Incorrectly Classified Instances	73.1156%
Naïve Bayes Cross Validation	
% Correctly Classified Instances	5.5276%
% Incorrectly Classified Instances	94.4724%
MultiLayerPerceptron Cross-Val	
% Correctly Classified Instances	13.1281%
% Incorrectly Classified Instances	86.8719%
AdaboostM1 Cross Validation	
% Correctly Classified Instances	18.0276%
% Incorrectly Classified Instances	81.9724%

Table 4.4 WEKA Results: IHME VA's

4.2.1 Discharge Summaries

The results from the baseline:

OneR predicted 62.5% correctly classified predicting on Hypertension on Prototype 1 and then on Cardiac Catheterization on Prototype 2. ZeroR and J-Rip produced the same results across all the prototypes achieving 50% of correctly classified instances. ZeroR predicting Pneumonia which was expected as this was the majority class.

Overall, from the remaining algorithms, J48, Naïve Bayes, MultliLayerPerceptron, Adaboost.M1 and LogisticR the most accurate results were output from Prototype 2. However, only Adaboost.M1 was able to offer a high degree of accuracy with the results exceeding the target outlined by Anker. Of all the algorithms LogisticR was the poorest performer, with less than a 50% performance.

Of the three diseases classes – Pneumonia, Coronary Artery Disease (CAD) and Chronic Obstructive Pulmonary Disease (COPD), the disease which was most successfully classified was Pneumonia.

With J48, Adaboost.M1 Pneumonia achieved 100% sensitivity for all prototypes. Naïve Bayes delivered a sensitivity of 87.5% on Prototype 1 and 100% on Prototype 2. Both MultiLayerPerceptron and Logistic R performed to a similar standard circa 62% sensitivity except in the case of Prototype 2 on MultiLayerPerceptron where it achieved 87.5%.

In evaluation there are a number of reasons why pneumonia achieved the most accurate predications. Firstly if you consider the overall data set 50% of the cases were cited as patients suffering from pneumonia. This will have been an advantage with some of the algorithms. Ideally it would have been better to have equal numbers of disease cases, but in this sample it was not achievable. Also with pneumonia the signs and symptoms were more distinct than the other two disease groups. Although again a small sample and the author has already advised that the sample size inhibits its results from a statistical perspective, this does concur with the research findings and indeed results on data driven algorithms used for verbal autopsy. In a number of the algorithms the classifier identifies that importance of chest x-ray and that this procedure was unique to the pneumonia cases. Also that the symptom plural effusion was common in pneumonia cases, none reported for CAD and only one report for COPD.

After Pneumonia, CAD was the next most successful classification, although overall the results were poor. Although in prototype 2, Adaboost.M1 delivered 100% classified instances for the remaining classifiers performed badly with results from only 33-67%. Obtaining no way near the benchmark required for accuracy in cause of death diagnosis. On examining the success of the Adaboost.M1 results, the classifier choose a decision stump as its method successfully identifying both the chest x-ray and cardiac catheterization as the key attributes which would deliver an accurate result. In consideration as to why most of the classifiers performed poorly the main reasons were found that there were overlaps with the signs and symptoms of CAD and COPD. For example many of the CAD and COPD patients shared the symptom of problematic blood pressure, had undertaken catheterization procedures or had experienced a myocardial infarction. As a result there is less clear water between these classes and the net result being that the classifiers make errors/mistakes. Another reason is that some of these algorithms are complex in makeup, e.g. MultiLayerPerceptron and this complexity only serves to add confusion into as result with a small data set. With a larger data set, this learning algorithm may have better performed.

In terms of the classifiers ability to predict COPD, J48, MultiLayerPerceptron, Naïve Bayes delivered a 0% sensitivity on all classifiers. Only Prototype 2 provided some success with Adaboost 100% sensitivity and LogisticR achieving 33.3%, very disappointing. Investigation in the poor performance was seen as sample size only 3 out of the 16 samples were COPD cases and also there is the overlap of symptom with the CAD cases.

In relation to prototype 3, when the discharge summary corpus had been successfully put through GATE what became very clear was the sheer volume of SNOMED-CT concepts annotated. In total there were an astonishing 9981!, although, the results that occurred were not as expected. Naively, the expectation was that all the concepts would be all medical words or phrases. The output of the results did deliver these successfully but it also delivered a multitude of other words which potentially were going to cause an issue with the classifiers. Examples of the additional concepts which were appearing as annotations were "date", "other", "seen", "started", numeric numbers and there were many, many more. Although when reading the discharge summaries these words were important for context and it was clear how they benefit a physician it very much increased the complexity from a computational perspective. Using one of the discharge summaries as an example, here is a short except. In colour are the highlighted SNOMED-CT concepts.

"Left side shows fibrofatty plaque, mostly flat in common carotid and scattered heterogeneous plaque at the bulb. V-p lung scan was performed on date [may 24 2007], which showed low probability of pe".

Although some of these words such as "plaque" are really useful for the classifier would have had only the following words as an input.

"Left side, plaque, flat in common, scattered, plaque, bulb. p, date, low".

A decision had to be made on the next course of action. The root cause of the issue was that the SNOMED-CT concept file was so granular, which viewed as a key benefit but now had created an annotation list which was now so full of noise it was potentially preventing the obtainment of any meaningful results. Thinking around the issue, the course of action chosen was to reduce the SNOMED-CT concept annotations by only including the most frequent medical terms. This would enable a prototype to be built and results obtained. The reason this decision was taken was that the author wanted to test if increasing the medical concepts from prototypes 1 and 2, which only contained 21 and 24 concepts (attributes) respectively, would increase the classifiers ability to predict cause of death in Prototype 3. If Prototype 3 production was aborted, this question would not be answered

Through undertaking a layman's assessment of the corpus 146 of the most frequent occurring medical terms were extracted and then the classifiers were built. The results of Prototype 3 proved to be very interesting.

Overall for Prototype 3, Adaboost.M1 delivered 100% sensitivity on all diseases and overall performed equal to prototype 2 on MultiLayerPerceptron and LogisticR, although overall this was not a good result as neither of these classifiers had performed well across the board of results. The worst performance was using Naïve Bayes only achieving 43.75%.

In conclusion, prototype 2 delivered the best results. Prototype 2 was either the best performing or equal best performer.   The conclusion drawn from this is that Prototype 2 benefited from the extra granularity with the terms for example “pulmonary hypertension” rather than hypertension, which gave a greater uniqueness between classes and therefore the classifier performed better. Prototype 1 came in second best followed by Prototype 3. Although this cannot be validated as another sample is not available for test, it is suspected that Prototype 3 would perform equal if not better to 1 and 2 had the data set been significantly larger. To explain with only 16 samples and 146 attributes there is too much sparseness of data values to produce an accurate result. Had there been 1600 samples the results may have been very different.

The results from Prototype 3 provide an opportunity for future NLP exploration to investigate if improvements could be made on a new prototype which would remove the majority of the noise.

4.2.2 Ghana Verbal Autopsy

The first prototype of the verbal autopsy sample suffered the same issues as Prototype 3 within the discharge summaries. From a corpus of 2658 tokens, 551 SNOMED-CT concepts were identified. Again the sample contained a considerable amount of noisy words which were unlikely to have added credibility to the classifier, words such as… “out”, “before” “related”, “month”, “seventh.”

So as Prototype 3 from the discharge summaries the most frequent concepts were extracted, in this 51 to ensure that some results could be obtained.

<u>The results from the baseline: Story of Illness</u>

Due to the size of the sample cross validation could not be performed on this sample. As a result the training function had to be used, see Appendix U for full details. The results are listed below but not considered statistically valid:

OneR, ZeroR, J-Rip all delivered the same results 40% correctly classified instances and 60% incorrectly classified.

In terms of the remaining classifiers, Naïve Bayes, MultiLayerPerceptron and LogisticR all delivered 100% correctly classified instances whilst J48 and Adaboost.M1 delivered 60%.

The results from the baseline: CSV file

The remaining prototype using the CSV had similar results, as again the training function could be applied to the classifier.

It is fair to say that these results are not statistically valid. However what was a very interesting learning area that came out of the prototype of real worth was the annotation observations

Grammatical issues and missing spelling within the Ghana verbal autopsies was very prevalent. To illustrate using just one of the verbal autopsies every line in the document except one contained spelling and/or grammatical errors. The document consisted of 427 tokens and within that there were 19 spelling mistakes both with medical and non medical words. As a result some of the signs and the symptoms were not annotated by the gazetteer in GATE. This caused a number of signs and symptoms to be omitted when they should have been recognised as SNOMED-CT concepts. A few examples being the misspelling of “dizziness” as “diziness”, “bulging fontened” when it should have read “bulging fontenelle” [which is a build up of fluid on the brain in new born babies something that needs to be identified in a verbal autopsy questionnaire]. Other examples were “jaundice” spelt as “jaudice” and “breathing” spelt as “breating.”

Also there were examples of important symptoms and signs which were not annotated by the gazetteer. For example a local term “anidane” which describes pain in the lower abdomen whilst in pregnancy. This term also appears as a particular question in the structured section of the questionnaire asking if “anidane” had occurred. Also within the structured questionnaire it asks if symptoms of “afare” and “afam” had been present. Both of these are Ghanaian terms, “afare” is being too thin or malnourished looking at birth and “afam” means extremely sick and about to die. The gazetteer would not recognise any of these terms. Neither did it recognise another local term which appeared in some of the other verbal autopsies obtained, the term “asram.” In Ghana, asram is the main serious illness (in local language terms) and most mentioned by care givers in relation to newborn illness or death. The symptoms are described as causing green veins on a baby’s body, continuous crying and growing lean [94]. The cause of this is said to be either passed to the baby through jealousy, bad spirits or the devil has taken over the baby. The Ghanese people believe that if this occurs in a baby there is nothing that can be done [94-95]. This one experience really did illustrate the issue of using particular terminological systems and also questions their practical use in certain situations. Omittance of these very important key terms is very much a drawback for computational approaches. If only the core terminological system detail is used and does not allow for adaptations based on country. In researching it was found that in countries in Africa local terminology is injected into data driven algorithms to ensure that vital information is not lost [91]. This is a very important observation and key learning area from the research conducted.

4.2.3 IHME Verbal Autopsy

To refresh, this was a csv driven classifier with no GATE process. In total there were 132 attributes (all anonymised), 1592 verbal autopsy cases and 32 possible causes of death again, all of which were anonymised.

The results from the baseline:

ZeroR achieved 11.6% correctly classified instances, 19% was achieved via OneR and a more positive but still extremely poor result of 27.6% on J-Rip.

In terms of the other classifier results there was a wide range of results from 5.5% -26.8%, see Appendix V for the full set of results. What was very clear was that the results were very poor and an investigation was carried out to determine why this was the case.

This was a particular challenge due to the heavy anonymisation of the data which meant that all that could be observed were numbers. However, when taking the cause of death data and placing it into a histogram format, see fig: 4.2, some interesting results became clear. There were a real disparate number of causes of death. For example within the 1592 VA's there were only 5 examples of "x4" cause of death compared to 186 of "x16" cause of death. Looking at the overall results the only causes of death which were accurately predicted to a 90%+ sensitivity were "x6" and "x16" see Appendix V.

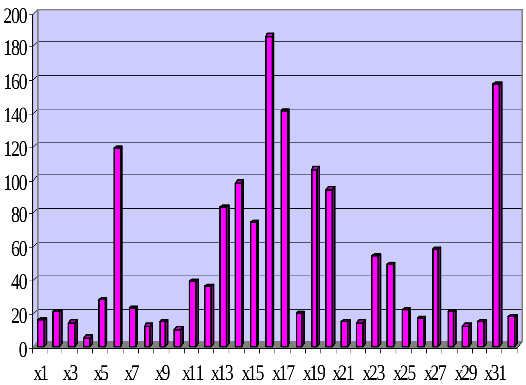


Fig 4.2: Cause of Death for the IHME sample

There could be a variety of reasons for this but given the limitations known about the data, two broad explanations are offered. Firstly, as both of these have a significant sample size this may have improved the results with the classifiers or that these causes of death have particularly distinct symptoms which enable the cause of death to be more accurately predicted.

Drawing the evaluation to nearly a to close, I would like to end with some general comments about the tools and the systems used.

4.3 Evaluation of SNOMED-CT

SNOMED-CT is clearly very granular, has good coverage and possesses a demonstrated clear ability to deal with composite phrases. For this task it is felt that this did impede the results from this particular prototype rather than enhance them. SNOMED-CT found it problematic to identify commercial names for medicines against their generic names. Possibly not a major issue in verbal autopsies, but it was very apparent with the US discharge summaries. Its ability to deal with certain words or phrases was an interesting observation – it annotated fever with chills but not fever with chill, it annotated "lung cancer" but not "cancer of the lung." Also it does not have diabetes or hypertension as a SNOMED-CT concept. Although I understand why, it provides the full preferred name which assists in standardisation and increases machine readability on patient notes. This means that these two terms which were seen regularly within the medical text documentation were not identified. These will not be lone examples; there will be others.

However, it is a multi purpose nomenclature and currently is not used for VA coding as ICD-10 is the recognised terminology. However, its composite phrasing was a benefit to the prototype. If the opportunity was present what would have been good would have been some expert medical advice on which concepts were needed to be included to assist in determining cause of death.

4.4 Evaluation of GATE

GATE is open source software and seemed to be slow at times on processing. It was difficult to master, and not intuitive to use. For an NLP novice, building the gazetteer and also the annotation pipeline required considerable thought and work to get right. The lack of output function into ARFF or csv was disappointing and caused integration issues to other tools, although this was overcome by the use of python. The lookup through the Gazetteer as a visual was very clear and easy to understand. Although what was disappointing again is that there is no integrated spell checker or plugin that can be attached. This would have certainly improved the annotation on the Ghana verbal autopsy sample.

4.5 Evaluation of WEKA

WEKA performed well with the prototype. It has a wealth of learning algorithms to choose from which enabled a wide range of them to be used with confidence and the results documented. The only criticism with WEKA is that with so much choice it is not clear which to use, and with such a broad functionality reading up on their purpose and descriptions is required before a final selection can be made.

4.6 Evaluation Feedback from VA Researchers:

I submitted a draft of this report to:

Betty Kirkwood, London School of Hygiene and Tropical Medicine.

Karen Edmonds, London School of Hygiene and Tropical Medicine.

Sammy Danso, Kintampo Health Research Centre, Ghana.

Dr. Abraham D Flaxman, Institute of Health Metrics and Evaluation, Washington University, USA.

Sean T Green, Institute of Health Metrics and Evaluation, Washington University, USA.

Saman Hina, Assistant Professor at NED University of Engineering & Technology, Karachi and currently a PhD Student at Leeds University.

Subsequent feedback received back to Student and Project Supervisor:

Dr Abraham D Flaxman (26th August 2010).

> "This looks really nice, just the kind of thing I was hoping our data could help with. I'm glad our data was helpful".

Sean T Green (29th August 2010).

> "I thought your project covered a lot of different aspects of VA thoroughly".

Saman Hina (25th August 2010).

> "You did great job to complete this project as the data used in this project is not simple at all and understanding the complexities of free text in natural language and data standards was really appreciable in this short duration of your project time".

Sammy Danso via Dr Eric Atwell whilst the project was being completed.

> "I have been following Rebecca's project on her blog and I must admit that I'm impressed with her progress made so far."

Betty Kirkwood was on annual leave at the time of project report completion.

Chapter 5: Conclusions

With a smaller data sample than desired, it is very hard to draw some exacting conclusions with regard to the specific results included in this project. The prototype itself although not handling large volumes of verbal autopsies for this particular data set clearly had the ability and robustness to process much larger samples and I am confident that if larger samples were available it would have delivered results where some more definite conclusions could be drawn. However, the building of this prototype and going step by step through the process has proved to be a very worthwhile and valuable exercise. It did enable all the requirements of the project to be met as there was an ability to examine, illustrate, understand and face the real challenges of this problem space. The findings do add to the existing research in this field.

What the model does illustrate is the sheer complexity of the task and the challenges that surround extracting information from medical documents such as verbal autopsies and discharge summaries when attempting to address with a computational approach.

What resonated with the samples obtained is despite being small, the uniqueness of each one. Every patient is different; all have a story/history which is unique to them. Trying to extract the information and then arrive at a cause of death is a difficult task for a physician let alone a computer. The importance of local knowledge and local context has proved to be crucial in the process. Local terminology is important, terms such “afam”, “atare” and “asram” are not included in international terminologies. These are important terms and should not be ignored. If someone from Northern Canada was complaining of fever chills and nausea and vomiting before they died you would not think they had malaria but you would if a person had those symptoms in Ghana! This is an extreme case but also the issue also occurs at the subtle level which was brought out in the results of the prototype; compounding the issue from a NLP and computational perspective.

To explain where there was clear water between the symptoms then the prototype/classifier performed best. This supports the view that a classifier will more accurately predict cause of death when symptoms are distinct. Even with a small sample the prototype more accurately classified the pneumonia cases than coronary artery disease and chronic obstructive pulmonary disease where similar signs and symptoms are often present.

This does support the view that one terminological system does not fit all. The project has demonstrated that SNOMED-CT worked well on the US discharge summaries, unsurprising a system developed in the US/UK. Although it performed less well given autopsy data from Africa, where culture and tradition have different words and even meanings for some diseases which are just not recognized in the western world.

The prototype also very pointedly illustrated the NLP issues and challenges when dealing with both free text and strucutured text formats. Although there were abbreviations in the discharge summaries often the SNOMED-CT gazetteer was able to identify these due to its ability to acknowledge synonyms which proved helpful in the identification of signs and symptoms. Abbreviations were much less of an issue with the verbal autopsies although with them the biggest issue along with local terminology was the high degree of misspelling which impacted on the ability to extract vital information. This is not surprising as the interviewers are often lay people and the "coders" although trained would not have the education of a physician. This is not an easy issue to address.

The WHO is pressing for standardisation of documentation and overall this does appear to be the right action path, as standardisation does increase the possibility of machine readability of documentation and would also enable more research to be compared and evaluated, something that is very much lacking in this problem space. However, the pace of change is slow and the process of moving to standardisation is fragmented. Increasing the machine readability could potentially reduce manual resources expended although the cost of set up and maintenance could very easily outstrip the cost of employing a coder for example in Ghana. Therefore not only is there a computational challenge to address but also cost issues, a major issue for developing countries.

Based on all the issues and challenges that have been presented and drawn out through this project, and the fact that to date they are still unresolved despite concerted efforts from various organisations and bodies all around the world it is unlikely that a computer will be able to take the role of both the coder and the physician to establish an accurate cause of death in the near future.

5.1 Future Work:

Research has shown that data sample sizes together with an associated gold standard is a major issue overall in this problem space. To be able to take this forward from a computational approach, larger samples need to be gathered and importantly conducted under the same protocols so that comparability can be assessed. Only then can computational processes start to move forward. Standardisation is also key so that machine learning becomes a viable option not only to assist in developing more accurate predictors of cause of death but also to assist with cost control.

Alternatives are needed to physician review as it is relatively cost ineffective and not feasible when assessing large numbers of questionnaires. More research needs to be carried out using the data driven methods of Logistic Regression, ANN and Bayesian approaches to provide a real alternative that can handle volume case load and predict with a high degree of accuracy and consistency cause of death.

In acknowledging the key benefits of the physician review and predefined expert algorithms, local knowledge, local custom awareness and experience, there may be an argument to look at how case based reasoning could assist in the process. Through case based reasoning a system would be developed to diagnose cause of death based on a series of typical cases. When conducting research for this project, case base reasoning was researched. In general case based reasoning can be very consistent if a standard system is developed. However, when undertaking the research only one reference was found within some documentation by the WHO that advised that currently (as of 2005) no such systems had been developed [91]. This may be a suitable area for research moving forward.

In final conclusion, data driven research may feedback into improved design of standardised questionnaires. If we have a better understanding of which features and questions are useful in automated diagnosis, this can inform the design of questionnaires, so that the VA can be simplified.

.

.

References:

[1] Edmond KM, Quigley MA, Zandoh C, et al. 2008. Aetiology of stillbirths and neonatal death in rural Ghana : implications for health programming in developing countries. *Paediatric and Perinatal Epidemiology* .**22**, pp.430-437.

[2] Edmond KM, Quigley MA, Zandoh C, et al. 2008. Diagnostic accuracy of verbal autopsies in ascertaining the causes of stillbirths and neonatal death in rural Ghana. *Paediatric and Perinatal Epidemiology*.**22**, pp.417-429.

[3] Byass P.2007. Who Needs Cause-of-Death Data*? PLoS Med*. **4**(11), pp. 1715-1716.

[4] Mathers CD, Ma Fat D, Inoue M, et al. 2005. Counting the dead and what they died from: an assessment of the global status of cause of death data. *Bulletin World Health Organization*. **83**(3), pp.171-180.

[5] Baiden F, Bawah A, Biai S, et al. 2007. Setting international standards for verbal autopsy. *Bulletin World Health Organization*. **85**(8), pp.570-571.

[6] Soleman N, Chandramohan D, Shibuya K. 2006. Verbal autopsy: current practices and challenges. *Bulletin of the World Health Organization*. **84**(3], pp.239-245.

[7] Biraud,Y. 1956. Méthodes pour l'enregistrement par des non médecins des causes élémentaires de décès dans les zones sous-développées. *Geneva: World Health Organization:* WHO document HS/60.

[8] Lay reporting of health information. 1978. *Geneva: World Health Organization.*

[9] Kielman AA, Dsweemer C, Parker R et al. 1983. Analysis of morbidity and mortality Vol:1 Integrated nutrition and health care. pp. 172-214. Baltimore: John Hopkins University Press.

[10] Garenne M, Fontaine O. 2006.Assessing probable causes of death using a standardized questionnaire: a study in rural Senegal. *Bulletin of the World Health Organization*]. **84**(3), pp.248-253.

[11] Fauveau V. 2006. Assessing probable causes of death without death registration or certificates: a new science? *Bulletin of the World Health Organization* .**84**(3), pp.246-247.

[12] Fottrell E, Byass P. 2010.Verbal Autopsy: Methods in Transition. *Epidemiol Review (advanced access publication 4 March)*, pp.1-18.

[13] Anker M, Black RE, Coldham C, et al. 1999. A standard verbal autopsy method for investigating causes of death in infants and children. Geneva: *World Health Organization.*

[14] King G, Lu Y. Shibuya K. 2010. Designing Verbal Autopsy Studies. *Population Health Metrics forthcoming*. Available from http://gking.harvard.edu/files/desva.pdf

[15] Boulle A, Chandramohan D, Weller P. 2001 A case study of using artificial neural networks for classifying cause of death from verbal autopsy**.** *International Journal of Epidemiology*. **30,**. pp.515-520.

[16] Quigley M, Chandramohan D, Rodgriges LC. 1999. Diagnostic accuracy of physician review, expert algorithms and data-derived algorithms in adult verbal autopsies. *International Journal of Epidemiology*. **28,** pp.1081-1087.

[17] Reeve, BC, Quigley, M.1997. A review of data-derived methods for assigning causes of death from verbal autopsy data. *International Journal of Epidemiology*. **26**(5), pp.1080-1089.

[18] De Keizer NF, Abu Hanna A, Zweersloot-Schonk JH. 2000. Understanding terminological systems I: Terminology and Typology. *Methods of Information in Medicine*. **39**(1), pp.16-21.

[19] Mori, RA, Consorti, F, Galeazzi, E. 1998. Standards to support development of terminological systems for healthcare telematics. . *Methods of Information in Medicine*.. **37**(4-5), pp.551-563.

[20] Lusignan De, S. 2005. Codes, classifications, terminologies and nomenclatures: definition, development and application in practice. *Informatics in Primary Care.* **13**(1), pp.65-69

[21] Cornet.R.2006. Methods for auditing medical terminological systems. Ph.D. thesis, Academic Medical Center, Universiteit van Amsterdam.

[22] Geller, Perl,Y, Cornet,R. 2009. Auditing of Terminologies. *Journal of Biomedical Informatics.* **42**(3), pp. 407-411.

[23] World Health Organization Website. 2010. *The History of ICD-10.* Available from: *http://www.who.int/classifications/icd/en/HistoryOfICD.pdf*

[24] World Health Organization Website. 2010. *WHO Webpage on ICD* . Available from:http://www.who.int/classifications/icd/en/

[25] World Health Organization Website. 2010. *Health Statistics and Health Information Systems*. Available from http://www.who.int/evidence/en/

[26] IHTSDO Website. 2010. *SNOMED-CT.* Available from: http://www.ihtsdo.org/snomed-ct/

[27] College Of American Pathologists Website .2010. *SNOMED-CT*. Available from: http://www.cap.org

[28] National Library of Medicine Website. 2010. *SNOMED-CT.*

Available from: http://www.nlm.nih.gov/index.html

[29] National Library of Medicine Website. 2010. *UMLS factsheet.* Available from: http://www.nlm.nih.gov/pubs/factsheets/umls.html

[30] Lindberg, DA., Humphreys, BL, McCray, AT. (1993). The Unified Medical Language System. *Methods Information in*. *Medicine*. **32**(4), pp.281-91.

[31] Kleinsorge, R, Willis J, Browne, A et al. 2006. AIMA Tutorial T12 Unified Medical Language System UMLS Overview Presentation. Library of Medicine National Institutes of Health U.S. Dept. of Health & Human Services. Available from: http://www.nlm.nih.gov/research/umls/pdf/AMIA_T12_2006_UMLS.pdf

[32] Vikstrom, A, Nystrom, M, AH, Feldt, H et al 2010. Views of diagnosis distribution in primary care in 2.5 million encounters in Stockholm: a comparison between ICD-10 and SNOMED-CT. *Inform Prim Care*.**18**(1), pp.17-29.

[33] Brown, SH, Elkin, PL, Bauer, BA, et al. 2006. SNOMED-CT®: Utility for a General Medical Evaluation Template. *AMIA Annual Symp Proc*. pp101–105.

[34] Brown SH, Rosenbloom, ST, BAUER, BA et al.2007. Direct Comparison of MEDCIN® and SNOMED-CT® for Representation of a General Medical Evaluation Template. *AMIA Annual Symp Proc*.pp75–79.

[35] Campbell JR, Carpenter, P, Sneiderman, C et al. 1997. Phase II evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions, and clarity. CPRI Work Group on Codes and Structures. *Journal American Medical Informatics Association*. May–June; 4(3), pp. 238–251.

[36] Rosenbloom ST, Brown SH, Froehling, D et al.2009 Using SNOMED-CT to represent two interface terminologies. *Journal American Medical Informatics Association*. Jan–Feb: 16(1),pp81–88**.**

[37] Campbell JR, Payne TH. 1994. A comparison of four schemes for codification of problem lists. *Proc Annu Symp Comput Appl Med Care*. pp.201-205.

[38] i2b2 Website.2010. Fourth i2b2/VA Shared-Task and Workshop: Challenges in Natural Language Processing for Clinical Available from: https://www.i2b2.org/NLP/Relations/

[39] Lussier YA, Bodenreider, O. 2007.Clinical Ontologies for discovery applications.In: BAKER CJO, CHEUNG KH, ed. 2007.Semantic Web: Revolutionizing knowledge discovery in the life sciences: Springer pp. 101-119.

[40] Taira RK, Soderland SG. 1999**.** A statistical natural language processor for medical reports. *Proc AMIA Symp*. pp.970-974

[41]Charlet J, Bachimon, B, Jaulent MC. 2006. Building medical ontologies by terminology extraction from texts: An experiment for the intensive care units *Computers in Biology and Medicine*. **36** (7-8),pp.857–870

[42] Haug PJ, Koehler S, Lau LM et al 1995. Experience with a mixed semantic/syntactic parser**.** *Proc Annu Symp Comput Appl Med Care* pp.284-288

[43] Hahn U, Romacker M, Schulz S. 2002. MEDSYNDIKATE: a natural language system for the extraction of medical information from findings reports. International Journal Medical Informatics. **67**(1-3), pp 63-74.

[44] University of Tokyo 2010. *Genia Tagger* and MEDIE applications. Available from: http://www.u-tokyo.ac.jp/index_e.html

[45] Aronson AR. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*. pp.17-21

[46] Friedman C, Shagina L, Lussier Y, et al. 2004. Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc*. **11**(5), pp.392-402.

[47] MEDLEE Website. 2010. MedLEE home page. Available from: http://zellig.cpmc.columbia.educ/medlee/demo.

[48] Bodenreider, O, Mitchell JA, McCray, AT. 2002. Evaluation of the UMLS as a terminology and knowledge resource for biomedical informatics. *Proc AMIA Symp*. pp.61-5.

[49] Morrison FP, Li Li MS, Lai A et al. 2009. Repurposing the Clinical Record: Can an Existing Natural Language Processing System De-identify Clinical Notes? *Am Med Inform Assoc.* 16(1), pp. 37–39.

[50] McCormick PJ, Elhadad N, Peter, D. 2008. Use of Semantic Features to Classify Patient Smoking Status. *AMIA Annu Symp Proc*. pp.450–454.

[51] Melton, GB, Hripcsak G. 2005. Automated Detection of Adverse Events Using Natural Language Processing of Discharge Summaries. *J Am Med Inform Assoc*. **12**(4), pp.448–457.

[52] Natural Language Applications Website. Press Release Pages. Available from http://www.nlpapplications.com/press10-06-08.html

[53] Lussier YA, Shagina L, Friedman C, 2001. Automating SNOMED coding using medical language understanding: A feasibility study. *Proc AMIA Symp.* pp 418–422.

[54] Fan WJ, Freidman, C. 2008. Semantic reclassification of the UMLS concepts. *Bioinformatics*. **24**(17), pp.1971-1973.

[55]Wilcox A, Hripcsak G.1998. Knowledge discovery and data mining to assist natural language understanding. *Proc AMIA Sym.* pp 835–839.

[56] Zenq TQ, Goryachev S, Weiss S, et al. 2006.Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system *BMC Medical Informatics and Decision Making* **6**(30).

[57] Friedman C, Alderson PO, Climino JJ et al. 1994. A general natural language processor for clinical radiology. *JAMAI*. **1**, pp161-174.

[58] Cunningham H. 2000. *Software Architecture for Language Engineering*. Ph.D. thesis, University of Sheffield.

[59]Cunningham H, Maynard D, Bontcheva K et al.2002. GATE: An architecture for the development of robust HLT applications. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics-Philadelphia, Pennsylvania.* pp 168-175.

[60] GATE Website.2010. *Gate Website*. Available from http://gate.ac.uk/

[61] Rapidminer Website.2010 *RapidMiner software*. Available from: http://rapid-i.com/content/view/181/190/

[62] University of Munich Website. 2010. *ELKI software*. Available from: http://www.dbs.ifi.lmu.de/research/KDD/ELKI/

[63]WEKA Website. 2010. *WEKA Website*. Available from: http://www.cs.waikato.ac.nz/~ml/weka/

[64] Hall M, Frank E, Holmes G et al. 2009. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter* .**11**(1), pp10-18.

[65] Shearer C. 2000. The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*. **5**(4), pp.13-22.

[66] CRISP-DM Website. 2010. *CRISP-DM Process*. Available from http://www.crisp-dm.org/Process/index.htm

[67] Mosby's Medical Dictionary (8^{th} edition). 2009. Mosby Elsevier.

[68] Measure DHS Project Support Website. 2010. *Measure DHS*. Available from: www.measuredhs.com

[69] Green ST, Flaxman AD. 2009. Machine Learning Methods for Verbal Autopsy in Developing Countries. *Association for the Advancement of Artificial Intelligence*. pp126-127.

[70] AAAI Spring Symposium on Artificial Intelligence for Development (AI-D) WEBSITE .2010. *Artificial Intelligence for Development: Stanford University 23-24 March 2010*. Available from:http://ai-d.org/program2010.html

[71] Long WL.2005. Extracting Diagnoses from Discharge Summaries. *AMAIA Symposium Proceeding*. pp.470-474.

[72] Chandramohan D, Maude GH, Rodgriges, et al. 1998. Verbal autopsy for adult deaths: their development and validation in a multicentre study. *Tropical Medicine and International Health*. **3**(6), pp.436-446.

[73] Anker M. 1997. The effect of misclassification error on reported cause-specific mortality fractions from verbal autopsy. International Journal of Epidemiology. **26**(5). pp.1090-1096.

[74] Maude GH, Ross DA. 1997. The effect of different sensitivity, specificity and cause specific mortality rates in children from studies using verbal autopsies. *International Journal Epidemiology*. **26**(5), pp.1097-1106.

[75] Chandramohan D, Setel P, Quigley M. 2001.Effect of misclassification of causes of death in verbal autopsy: can it be adjusted? *International Journal Epidemiology*. **30**(3), pp.509-514.

[76] Kalter HD, Gray RH, Black, R et al. 1990. Validation of post mortem interviews to ascertain selected cause of death in children. *International Journal of Epidemiology*. **19**, pp.380-386.

[77] Snow, R, Armstrong JRM, Forster D, et al. 1992. Childhood deaths in Africa: uses and limitations of verbal autopsies. *Lancet.* **340,** pp.351-355.

[78] Mobley,CC, Boerma, JT, Tituss, S et al. 1996. Validation study of verbal autopsy method for causes of childhood mortality in Namibia. *Journal of Tropical Pediatrics*. **42**. pp.365-369.

[79] Murray C, Lopez A, Feehan D, et al. 2007. Validation of the Symptom Pattern Method for Analyzing Verbal Autopsy Data. *PLos Medicine.* **4**(11), pp.1740-1753.

[80] Quigley MA, Amstrong Schellenburg JRM, Snow RW.1996. Algorithms for verbal autopsies: validation study in Kenyan children. *Bulletin World Health Organization*. **74**(2), pp.147-154.

[81] Chandramohan D, Rodgriges LC, Maude GH, et al. 1998. The validity of verbal autopsies for assessing the causes of institutional maternal death. *Studies in Family Planning*. **29**(6), pp.414-422.

[82] Byass P, Fottrell E, Lan Huong, D et al. 2006. Refining a probabilistic model for interpreting verbal autopsy data. *Scandinavian Journal of Public Health* **34**(1), pp.26-31.

[83] Quigley MA, Chandramohan D, Rodgriges LC. 2000. Validity of data-derived algorithms for ascertaining causes of adult death in two African sites using verbal autopsy. *Tropical Medicine and International Health*. **5**(1), pp.33-39.

[84] Fottrell E, Byass P, Ouedraogo TW et al. 2007. Revealing the burden of maternal mortality: a probabilistic model for determining pregnancy-related causes of death from verbal autopsies**.** *Population Health Metrics*. **5**(1).

[85] Byass, P, Huong DL, Minh HV. 2003. A probabilistic approach to interpreting verbal autopsies: methodology and preliminary validation in Vietnam. *Scandinavian Journal of Public Health Supplement.* **31** (Supplement 62) pp.32-37.

[86] Biruk T, Georges R, Tekebash A, Tolcha K. 2008. Evaluating the performance of the InterVA Model for determining AIDS mortality in the adult population of Addis Ababa. *European Population Conference: Barcelona: Spain.*

[87] Fantahun, M, Fotrell, Berhane Y et al. 2006. Assessing a new approach to verbal autopsy interpretation in a rural Ethiopian community: the InterVA model. *Bulletin World Health Organization.* **84**(3), pp.204-210.

[88] King, G. 2010. *Gary King website – VA: Software for Analyzing Verbal Autopsy Data.* Available from: http://gking.harvard.edu/va/

[89] King, G, Lu, Y. 2008. Verbal autopsy methods with multiple causes of death. *Statistical Science* [online]. **23**(1), pp. 78-91.

[90] Witten IH, Frank E.ed. 2005. Data mining: Practical machine learning tools and techniques. San Francisco: Elsevier.

[91] WHO Technical Consultation on Verbal Autopsy Tools. 2004. WHO Conference, Tailores, France.

[92] Altman DG, Bland MJ.1994. Diagnostic Tests 1: Sensitivity and Specificity. *British Medical Journal*. 308, pp.1552

[93] Free Dictionary. Available from https://Encyclopedia.thefreedictionary.com/sensivitity

[94] Bazzano1 AN, Kirkwood BR, Tawiah-Agyemang C et al. 2008. Beyond symptom recognition: care-seeking for ill newborns in rural Ghana. *Tropical Medicine and International Health*. 13(1). pp. 123–128.

[95] West,R. 2010. Email to Karen Edmond, 15 August.

Appendix A: Reflection on the Project

Acquiring a suitable data set for this project was extremely challenging, but I guess you have picked this up reading the project report. The learning point here is that if you don't have the data at project commencement or a guarantee that you will be provided with the data (coming from a known and trusted source) my advice would be to think long and hard as to whether you move forward with your project idea. Despite me being very well organised, with a good project plan, the lack of data put considerable strain and injected needless worry into the project. Although I dealt with it and overcame the challenge, it does add to an already demanding experience.

To explain, for all the MSc students reading this, the period between Christmas vacation and September is a long one. You have your exams in Feb so you are revising for them; you finish and then get straight into Semester 2. At that point you are working on project ideas and starting your literature search for your project. Then before you know it your May exams have arrived and by this time you should have got the bulk of your literature review done. It is a major juggling act and by this time you are very tired and there are still 4 months to go. So please learn from my experience, if you can't guarantee 100% that you can get your hands on the data then think twice about going down this route.

Another thought is make sure you choose something that ideally you are interested in or passionate about. I know that sounds a really obvious thing to do but it isn't. My experience is some students are so caught up with Semester 2 and exams that sometimes the project is an afterthought and then once they start it isn't what they think. So in the January have a think about what you are going to do and ensure that if a project intrigues you then make an appointment to see the lecturer concerned and ask them about the project, make sure you understand the subject and also assess whether you are going to find it engaging. In terms of my own experience, I was fascinated with this subject and when times got hard it was this fact alone that enabled me to keep my enthusiasm, drive and determination to succeed.

Planning your project cannot be underestimated. A good project plan put up somewhere in you home helps. Don't just make one and then file it, placing it somewhere prominent helps you to keep track of where you are and where you are going.– hopefully going forward! I also set up a blog. At first it seemed really strange documenting what I had done or was going to do on a blog (I'm not an active blogger) however I found real benefits in doing so. I put all my key literature review on it, my project plan and also any thoughts that I had including challenges and successes. When I had to write my report the blog helped me and it also gave visibility to my project supervisor.

Now I have finished, I recognize how important organization was to my project. If you are undertaking a project in a subject area unknown to you, which mine was, in the medical domain do allow yourself time to understand the terminology associated with the subject. You are adding an additional dimension to your project alongside the computing. My advice is when assessing the time to understand a new subject assess it then double it. This will ensure that you give yourself enough time to do the subject matter justice and allow yourself enough time to absorb it, something that cannot be underestimated. When you are conducting your literature review you will find yourself with an extraordinary amount of papers. I found the best way of managing them was to put them into piles based on research area/angle and then label them up with a highlighted marker – A,B,C, etc with the date they were produced. This saved me time when I was trying to my find papers and quickly provided me a chronological view of my research for my chapter 1.

Also I would say that Semester 3 is a very different from the previous semesters. During Semester 1 and 2 you spend a considerable amount of time with your class mates, sharing knowledge experiences and helping each other out where you can. When Semester 3 comes you stop seeing your class mates so much as there are no lectures to attend, it's just you and the project. It is tempting to stay at home and do your project but I found it beneficial to come to the University and meet up with class mates at least once a week. At times your class mates are your best stress relievers as they are going through the same thing as you especially in late July/August. At that point you are well entrenched within your project but at the same time have feelings about whether you are ever going to complete it! Completely irrational, but it does go through your mind.

Do see your project supervisor every week, you can solicit feedback and it's an opportunity to discuss any problems and issues. When I was doing the second part of the project building the classifiers I only had experience of "decision trees." I recognised that I needed a much broader knowledge of machine learning than I had acquired in lectures. So the ability to read up on subject material and then go and speak to my supervisor to check that my understanding was sound was an important and valuable resource. Another opportunity to gauge how you are progressing is to present your project to both your assessor and superviser. Get yourself prepared on the structure of the meeting (I used a powerpoint presentation which worked well for me) to guide the proceedings. It's a great opportunity to gain feedback from the assessor. I got some good advice on base lining in WEKA and based on feedback I wrote a python program which helped to address an implementation aspect of my project.

Finally, I would say that this has been the most challenging academic task that I have ever done, although overall I have found it to be a rewarding and interesting experience. After being out of education for 16 years it was a challenge to return to academic study. Although there have been times when I was stretched to the limits, completing the project really has consolidated my learning from the whole year and I feel that I am better placed for the new challenges to come.

Appendix B: Interim Report

Supervisor's comments on the Interim Report

A detailed and clear statement of the background to the problem, accessible to computer scientists (as well as medics). I noticed a few typos, corrected in the report; otherwise, excellent English. Impressively thorough use of references to back up your explanation of the background - reads like a survey journal paper.

CRISP-DM seems sensible for overall methodology, but your plans are lacking in detail as to what exactly you will have to do at each stage; in particular, what sort of "Models" you will hope to come up with, and how exactly they will be evaluated. These details should become clearer as the project progresses, but always remember ~~that~~ Evaluation contributes 20% of your mark so you will need a systematic approach to this.

It is not clear how "hard" or challenging this project is in terms of Computing work - as you note, most previous work is in Medical sources. In your Final Report, you must make clear the computing challenges you faced (and overcame, for the most part, I hope!) so we can give you full credit.

Overall, Impressive write-up so far, but Computing design and implementation need to start a.s.a.p.

[signature] 22/02/2016

Appendix B: Interim Report (continued)

Assessor's comments on the Interim Report

Clear Objectives and an interesting topic.
Very good written English.
However, minimum requirements must be measurable and should not include general objectives such as "Understand the purpose of VAs". Although Crisp-DM looks a reasonable choice, you must also discuss alternative methods to justify your final choice.
A copy of the work plan should have been included.
The literature review is thorough on the medical side but has almost no computational content. Although there might be few mixed approaches, these need more detailed discussion and related work in text classification should be discussed as well.
With regard to evaluation + test data, there are no clear plans yet in place, which is worrying. Also, no practical work seems to have been done yet, which makes you behind in your project.

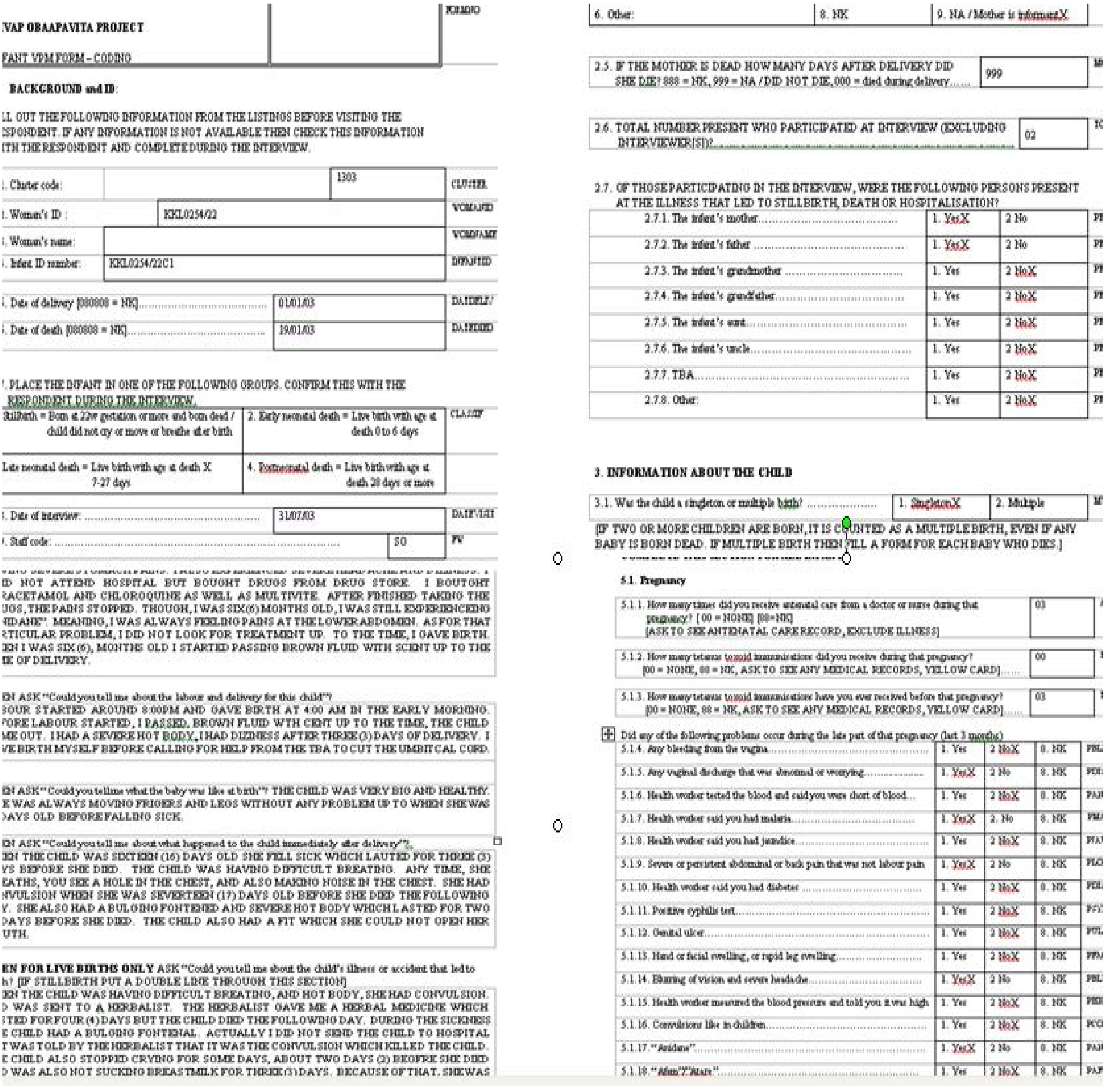
K. Mark

Appendix C: A Map of Countries where Verbal Autopsies are used

World map of countries (grey shading) where verbal autopsy methods are applied.
Source: Fottrell/Byass.2010. Methods in Transition
http://epirev.oxfordjournals.org/cgi/reprint/mxq003v1

Appendix D: Sample Ghana Verbal Autopsy

The verbal autopsy for the Ghana Neonates death is over 18 pages long. The screen shot below shows 4 of these pages. And clearly shows that the document contains but structured and non structured aspects.



VAP OBAAPAVITA PROJECT

FANT VPM FORM – CODING

BACKGROUND and ID:

LL OUT THE FOLLOWING INFORMATION FROM THE LISTINGS BEFORE VISITING THE SPONDENT. IF ANY INFORMATION IS NOT AVAILABLE THEN CHECK THIS INFORMATION ITH THE RESPONDENT AND COMPLETE DURING THE INTERVIEW.

Cluster code:	1303	CLUSTER
Woman's ID:	KKL0254/22	WOMANID
Woman's name:		WOMNAME
Infant ID number:	KKL0254/22C1	INFANTID
Date of delivery (080808 = NK)	01/01/03	DATDELI
Date of death (080808 = NK)	19/01/03	DATDDD

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

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

Date of interview:	31/07/03	DATEVISIT
Staff code:	50	FW

D NOT ATTEND HOSPITAL BUT BOUGHT DRUGS FROM DRUG STORE. I BOUGHT PARACETAMOL AND CHLOROQUINE AS WELL AS MULTIVITE. AFTER FINISHED TAKING THE DRUGS, THE PAINS STOPPED. THOUGH, I WAS SIX(6) MONTHS OLD, I WAS STILL EXPERIENCING "DANE". MEANING, I WAS ALWAYS FEELING PAINS AT THE LOWER ABDOMEN. AS FOR THAT PARTICULAR PROBLEM, I DID NOT LOOK FOR TREATMENT UP. TO THE TIME, I GAVE BIRTH. WHEN I WAS SIX (6), MONTHS OLD I STARTED PASSING BROWN FLUID WITH SCENT UP TO THE TIME OF DELIVERY.

EN ASK "Could you tell me about the labour and delivery for this child"?
OUR STARTED AROUND 8:00PM AND GAVE BIRTH AT 4:00 AM IN THE EARLY MORNING. BEFORE LABOUR STARTED, I PASSED, BROWN FLUID WITH CENT UP TO THE TIME, THE CHILD CAME OUT. I HAD A SEVERE HOT BODY, I HAD DIZINESS AFTER THREE (3) DAYS OF DELIVERY. I GAVE BIRTH MYSELF BEFORE CALLING FOR HELP FROM THE TBA TO CUT THE UMBILICAL CORD.

EN ASK "Could you tell me what the baby was like at birth"? THE CHILD WAS VERY BIG AND HEALTHY. SHE WAS ALWAYS MOVING FINGERS AND LEGS WITHOUT ANY PROBLEM UP TO WHEN SHE WAS DAYS OLD BEFORE FALLING SICK.

EN ASK "Could you tell me about what happened to the child immediately after delivery"?
WHEN THE CHILD WAS SIXTEEN (16) DAYS OLD SHE FELL SICK WHICH LAUTED FOR THREE (3) DAYS BEFORE SHE DIED. THE CHILD WAS HAVING DIFFICULT BREATHING. ANY TIME, SHE BREATHS, YOU SEE A HOLE IN THE CHEST, AND ALSO MAKING NOISE IN THE CHEST. SHE HAD CONVULSION WHEN SHE WAS SEVERTEEN (17) DAYS OLD BEFORE SHE DIED THE FOLLOWING DAY. SHE ALSO HAD A BULGING FONTENED AND SEVERE HOT BODY WHICH LASTED FOR TWO DAYS BEFORE SHE DIED. THE CHILD ALSO HAD A FIT WHICH SHE COULD NOT OPEN HER MOUTH.

EN FOR LIVE BIRTHS ONLY ASK "Could you tell me about the child's illness or accident that led to death? (IF STILLBIRTH PUT A DOUBLE LINE THROUGH THIS SECTION)
WHEN THE CHILD WAS HAVING DIFFICULT BREATING, AND HOT BODY, SHE HAD CONVULSION. I WAS SENT TO A HERBALIST. THE HERBALIST GAVE ME A HERBAL MEDICINE WHICH LASTED FOR FOUR (4) DAYS BUT THE CHILD DIED THE FOLLOWING DAY. DURING THE SICKNESS THE CHILD HAD A BULGING FONTENAL. ACTUALLY I DID NOT SEND THE CHILD TO HOSPITAL. I WAS TOLD BY THE HERBALIST THAT IT WAS THE CONVULSION WHICH KILLED THE CHILD. THE CHILD ALSO STOPPED CRYING FOR SOME DAYS, ABOUT TWO DAYS (2) BEOFRE SHE DIED AND WAS ALSO NOT SUCKING BREASTMILK FOR THREE (3) DAYS. BECAUSE OF THAT, SHE WAS

6. Other:	8. NK	9. NA / Mother is informantX

2.5. IF THE MOTHER IS DEAD HOW MANY DAYS AFTER DELIVERY DID SHE DIE? 888 = NK, 999 = NA / DID NOT DIE, 000 = died during delivery	999
2.6. TOTAL NUMBER PRESENT WHO PARTICIPATED AT INTERVIEW (EXCLUDING INTERVIEWER(S))?	02

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

2.7.1. The infant's mother	1. YesX	2 No
2.7.2. The infant's father	1. YesX	2 No
2.7.3. The infant's grandmother	1. Yes	2 NoX
2.7.4. The infant's grandfather	1. Yes	2 NoX
2.7.5. The infant's aunt	1. Yes	2 NoX
2.7.6. The infant's uncle	1. Yes	2 NoX
2.7.7. TBA	1. Yes	2 NoX
2.7.8. Other:	1. Yes	2 NoX

3. INFORMATION ABOUT THE CHILD

3.1. Was the child a singleton or multiple birth?	1. SingletonX	2. 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.)

5.1. Pregnancy

5.1.1. How many times did you receive antenatal care from a doctor or nurse during that pregnancy? [00 = NONE] [88=NK] [ASK TO SEE ANTENATAL CARE RECORD, EXCLUDE ILLNESS]	03
5.1.2. How many tetanus toxoid immunisations did you receive during that pregnancy? [00 = NONE, 88 = NK, ASK TO SEE ANY MEDICAL RECORDS, YELLOW CARD]	00
5.1.3. How many tetanus toxoid immunisations have you ever received before that pregnancy? [00 = NONE, 88 = NK, ASK TO SEE ANY MEDICAL RECORDS, YELLOW CARD]	03

Did any of the following problems occur during the late part of that pregnancy (last 3 months)

5.1.4. Any bleeding from the vagina	1. Yes	2 NoX	8. NK
5.1.5. Any vaginal discharge that was abnormal or worrying	1. YesX	2 No	8. NK
5.1.6. Health worker tested the blood and said you were short of blood	1. Yes	2 NoX	8. NK
5.1.7. Health worker said you had malaria	1. YesX	2. No	8. NK
5.1.8. Health worker said you had jaundice	1. Yes	2 NoX	8. NK
5.1.9. Severe or persistent abdominal or back pain that was not labour pain	1. YesX	2 No	8. NK
5.1.10. Health worker said you had diabetes	1. Yes	2 NoX	8. NK
5.1.11. Positive syphilis test	1. Yes	2 NoX	8. NK
5.1.12. Genital ulcer	1. Yes	2 NoX	8. NK
5.1.13. Hand or facial swelling, or rapid leg swelling	1. Yes	2 NoX	8. NK
5.1.14. Blurring of vision and severe headache	1. YesX	2 No	8. NK
5.1.15. Health worker measured the blood pressure and told you it was high	1. Yes	2 NoX	8. NK
5.1.16. Convulsions like in children	1. Yes	2 NoX	8. NK
5.1.17. "Asidane"	1. YesX	2 No	8. NK
5.1.18. "Afam"/"Atare"	1. Yes	2 NoX	8. NK

Appendix E: ICD-10 Chapters

Chapter	Blocks	Title
I	A00-B99	Certain infectious and parasitic diseases
II	C00-D48	Neoplasm's
III	D50-D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	E00-E90	Endocrine, nutritional and metabolic diseases
V	F00-F99	Mental and Behavioural disorders
VI	G00-G99	Diseases of the nervous system
VII	H00-H59	Diseases of the eye and addenda
VIII	H60-H95	Diseases of the ear and mastoid process
IX	I00-I99	Diseases of the circulatory system
X	J00-J99	Diseases of the respiratory system
XI	K00-K93	Diseases of the digestive system
XII	L00-L99	Diseases of the skin and subcutaneous tissue
XIII	M00-M99	Diseases of the musculoskeletal system and connective tissue
XIV	N00-N99	Diseases of the genitourinary system
XV	O00-O99	Pregnancy, childbirth and the puerperium
XVI	P00-P96	Certain conditions originating in the perinatal period
XVII	Q00-Q99	Congenital malformations, deformations and chromosomal abnormalities
XVIII	R00-R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX	S00-T98	Injury, poisoning and certain other consequences of external causes
XX	V01-Y98	External causes of morbidity and mortality
XXI	Z00-Z99	Factors influencing health status and contact with health services
XXII	U00-U99	Codes for special purposes

Source: WHO website: ICD-10 Coding Chapters

http://apps.who.int/classifications/apps/icd/icd10online/

Appendix F: Example of the Structure of a SNOMED-CT Concept

An example of the structure of a SNOMED CT concept

Concept:

- ConceptID: 22298006
- Fully specified name: myocardial infarction (disorder)

Descriptions:

- Preferred term: myocardial infarction
- Synonym: cardiac infarction
- Synonym: heart attack
- Synonym: infarction of heart

Relationships:

- **Defining relationships (is a)**
  - Concept: structural disorder of heart
    - Associated morphology: Infarct
    - Finding site: myocardium structure
  - Concept: injury of anatomical site
    - Associated morphology: infarct
    - Finding site: myocardium structure
  - Concept: myocardial disease
    - Associated morphology: infarct
    - Finding site: myocardium structure
- **Allowable qualifiers**
  - Qualifier: onset
  - Qualifier: severity
  - Qualifier: episodicity
  - Qualifier: course

Example of the structure of a SNOMED-CT concept

Source: Connecting for Health Website

http://www.connectingforhealth.nhs.uk/systemsandservices/data/snomed/snomed-ct.pdf

Appendix G. NLP Medical Text Analysis and Extraction Resources List

1. **BANNER.** Leaman, R., & Gonzalez, G. (2008). BANNER: an executable survey of advances in biomedical named entity recognition. In Pac Symp Biocomput (Vol. 652, p. 63).http://banner.sourceforge.net/
2. **Berkley Parser.** Petrov S, Barrett L, Thibaux R, and Klein D. 2006 Learning Accurate, Compact, and Interpretable Tree Annotation. In: COLING-ACL, 2006.Petrov S, and Klein D. Improved Inference for Unlexicalized Parsing. In: HLT-NAACL, 2007http://code.google.com/p/berkeleyparser/
3. **Bioscope Corpus.** Vincze V, Szarvas G, Farkas R, Móra G, and Csirik J. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In: BMC Bioinformatics 2008, 9(11) http://www.inf.u-szeged.hu/rgai/bioscope
4. **BIOSimplify.** Jonnalagadda, S., & Gonzalez, G. (2009). Sentence Simplification Aids Protein-Protein Interaction Extraction. In Languages in Biology and Medicine http://sourceforge.net/projects/biosimplify
5. **CCG Parser** Laura Rimell and Stephen Clark: Porting a Lexicalized-Grammar Parserto the Biomedical Domain. Journal of Biomedical Informatics, 2009.http://svn.ask.it.usyd.edu.au/trac/candc/
6. **ClearTK** Philip V. Ogren and Philipp G. Wetzler and Steven Bethard A UIMA toolkit for statistical natural language processing, UIMA for NLP workshop at Language Resources and Evaluation Conference (LREC) http://code.google.com/p/cleartk/
7. **cTakes**. https://cabigkc.nci.nih.gov/Vocab/KC/index.php/OHNLP_Documentation_and_Downloads
8. **DrugBank**. A knowledgebase for drugs, drug actions and drug targets. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M.Nucleic Acids Res. 2008 Jan;36(Database issue):D901-6. Epub 2007 Nov 29. http://drugbank.ca/
9. **dTagger** Divita G, Browne AC, Loane R. dTagger 2006. A POS Tagger. Proceedings of > AMIA Symposium. pp200-203. http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/dTagger/dtagger/doc/d...
10. **ENJU.** Yusuke Miyao and Jun'ichi Tsujii. 2008. Feature Forest Models for Probabilistic HPSG Parsing. Computational Linguistics. 34(1). pp. 35--80, MIT Press. http://www-tsujii.is.s.u-tokyo.ac.jp/enju/
11. **GATE.** http://gate.ac.uk/
12. **Genia Tagger.** Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii, Developing a RobustPart-of-Speech Tagger for Biomedical Text, Advances in Informatics -10th Panhellenic Conference on Informatics, LNCS 3746, pp. 382-392,2005. http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/
13. **MALLET.** McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." 2002. http://mallet.cs.umass.edu
14. **MedEx** Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. 2010. MedEx: a medication information extraction system for clinical narratives.J Am Med Inform Assoc. 17(1) pp.19-24.
15. **MEDIE.** http://www-tsujii.is.s.u-tokyo.ac.jp/MEDIE
16. **MEDLEE.** http://zellig.cpmc.columbia.edu/medlee/
17. **MedRA.** R. Fescharek, J. Kübler, et al. 2004. Medical dictionary forregulatory activities (MedDRA): Data

retrieval and presentation.International Journal of Pharmaceutical Medicine 18(5):259-269.http://www.meddramsso.com/

18. **MEDSYNDIKATE.** Hahn, U, Romacker M, Schulz S. 2002. MEDSYNDIKATE: a natural language system for the extraction of medical information from findings reports. International Journal Medical Informatics. 67(1-3), pp 63-74.

19. **MeSH vocabularies.** http://www.ncbi.nlm.nih.gov/mesh

20. **MetaMap and MetaMap Transfer.** Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. http://mmtx.nlm.nih.gov/

21. **MOBY** http://icon.shef.ac.uk/Moby/

22. **Natural Language Toolkit** - Garrette D, and Klein E. 2009. An Extensible Toolkit for Computational Semantics. In: Proceedings of the Eighth International Conference on Computational Semantics,Tilburg University, Netherlands, January. http://www.nltk.org/

23. **NegEX/ConText.** Chapman,W, Chu D, Dowling JN. "ConText: An algorithm for identifying contextualfeatures from clinical text" 2007. http://www.dbmi.pitt.edu/chapman/ConText.html

24. **OpenNLP.** http://opennlp.sourceforge.net/

25. **Python.** http://www.python.org/

26. **SimFind.** Jonnalagadda, S., Leaman, R., Cohen, T., & Gonzalez, G. (2010). A Distributional Semantics Approach to Simultaneous Recognition of MultipleClasses of Named Entities. In LNCS 6008. Presented at the CICLing. URL: http://www.public.asu.edu/~sjonnal3/SV_NER_src.zip

27. **SNOMED-CT.** http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/Publications/doc_UserGuide_Current-en-US_INT_20100131.pdf

28. **Specialist Lexicon**. http://lexsrv3.nlm.nih.gov/Specialist/Home/index.html

29. **Stanford Parser**. Klein D, and Manning CD. Fast Exact Inference with a Factored Modelfor Natural Language Parsing. Advances in Neural Information Processing Systems 15 (NIPS 2002), Cambridge, MA: MIT Press: 3-10 http://nlp.stanford.edu/software/lex-parser.shtml

30. **SYNTXT.** Haug PJ, Koehler S, Lau LM, Wang P, Rocha R, Huff SM. 1995 Experience with a mixed semantic /syntactic parser. Proc Annu Symp Comput Appl Med Care. pp. 284-8.

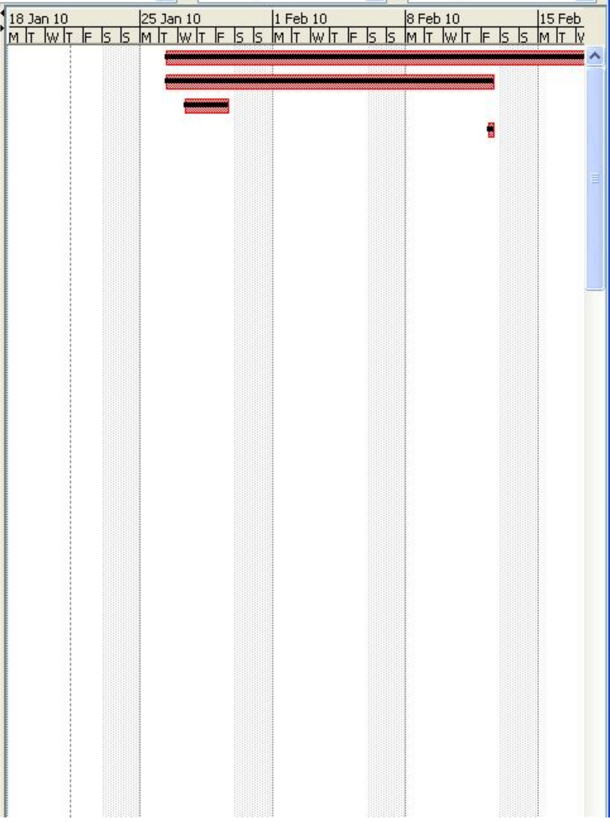
31. **UCLA Medical Imaging Informatics Toolkit.** http://www.mii.ucla.edu/nlp/

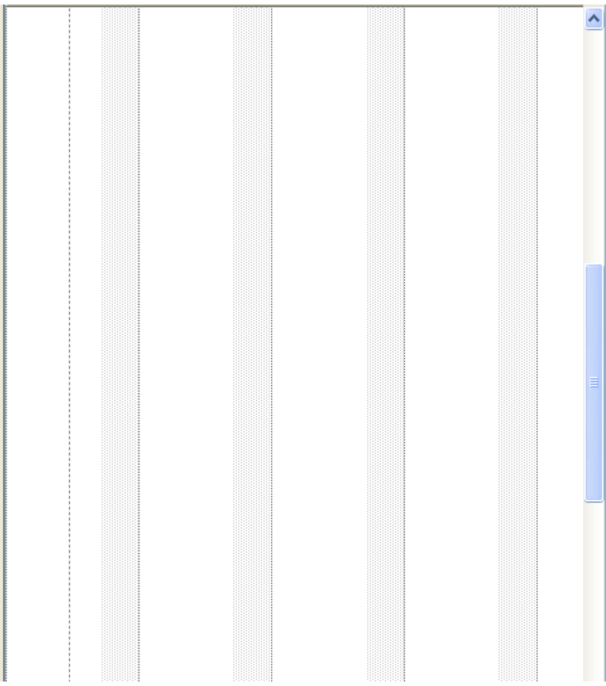
32. **UMLS vocabularies**. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods of Information in Medicine. 1993; 32(4):281-91. http://www.nlm.nih.gov/research/umls/

33. **WordNet.** Christiane Fellbaum and Joachim Grabowski and Shari Landes. Performance and Confidence in a Semantic Annotation Task. WordNet: an electronic lexical database. Chap. 9. p. 216--237. The MIT Press.1998. Ed. Christiane Fellbaum. Language, Speech and Communication. Cambridge, Massachusetts. http://wordnet.princeton.edu/

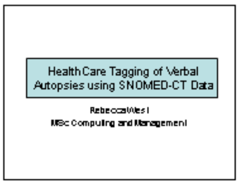
Appendix H: Project Plan

Name	Duration	S
PLANNING	43.875 days?	26/01/10 09:(
Project Ideas	13.875 days?	26/01/10 09:(
Initial Literature Search	2.875 days?	27/01/10 09:(
Selection Project	0.875 days?	12/02/10 09:(
Project Supervisor Allocated	0.875 days?	24/02/10 09:(
First Meeting with Supervisor	1 day?	03/03/10 09:(
Agree Weekly Meeting Appointment	1 day?	03/03/10 09:(
Initial Project Plan Draft	1 day?	08/03/10 08:(
Research Methodology Meeting	1 day?	08/03/10 09:(
Meeting with Supervisor	1 day?	10/03/10 09:(
Aims and Min Requirements Document Produced V1	1 day?	10/03/10 09:(
Aims and Min Requirements Document V2	1 day?	11/03/10 09:(
Aims and Min Requirements Document V3 submitted	0.875 days?	12/03/10 09:(
Natural Language Toolkit Introduction	0 days?	22/03/10 08:(
Easter Vacation (Exam revision)	17.875 days?	22/03/10 09:(
Blog set up (http://mscgirl.wordpress.com)	1 day?	23/03/10 09:(
Project Plan Redrafted	2.875 days?	24/03/10 09:(
WEKA - Introduction to and Sample Worksheets	4 days?	30/03/10 08:(
Semester 2 Work and Coursework	15 days?	19/04/10 08:(
Revision Week - WEKA (2 days)	5 days?	10/05/10 08:(
Semester 2 Exams	10 days?	17/05/10 08:(
NIH Web Based Course Completed	1 day?	18/05/10 08:(
SITUATION EVALUATION	10 days?	07/06/10 08:(
Supervisor Meeting	1 day?	02/06/10 08:(
DATA COLLECTION	11 days?	03/06/10 08:(
Literature Search II	11 days?	03/06/10 08:(
Work on Production of Interim Report	11 days?	03/06/10 08:(
Supervisor Meeting	1 day?	09/06/10 08:(
GATE Workshop	0.625 days?	10/06/10 11:(
Acquistion of initial medical texts for examination	1 day?	10/06/10 08:(
Review medical text with data mining tools	2 days?	10/06/10 08:(
Supervisor Meeting	1 day?	16/06/10 08:(
DEADLINE: Interim Report Due!!!	1 day?	18/06/10 08:(
SOLUTION DEFINITION PHASE	12 days?	21/06/10 08:(
Solution Definition Stage I - Scoping Prototype	2 days?	21/06/10 08:(
Supervisor Meeting	1 day?	23/06/10 08:(
Student Away	3 days?	24/06/10 08:(
Solution Definition Stage II - Build Protoype	3 days?	28/06/10 08:(
Project Write Up	4 days?	29/06/10 08:(
Feedback from Interim Report	1 day?	30/06/10 08:(
Supervisor Meeting	1 day?	30/06/10 08:(
Solution Defintion Stage III - Build and Revise Proto	4 days?	01/07/10 08:(
Project Write Up	3 days?	05/07/10 08:(
Supervisor Meeting - Initial Evaluation	1 day?	07/07/10 08:(
IMPLEMENTATION OF SOLUTION	10 days?	08/07/10 08:(
Implementation/ Evaluation of Solution (Stage I)	4 days?	08/07/10 08:(
Project Write Up	4 days?	12/07/10 08:(
Supervisor Meeting - 2nd Evaluation/Feedback	1 day?	14/07/10 08:(
Implementation/Evaluation of Solution (Stage II)	4 days?	15/07/10 08:(
Project Write Up	3 days?	19/07/10 08:(
Supervisor Meeting - 3rd Evaluation/Feedback	1 day?	21/07/10 08:(
Supervisor Meeting - Agree in Principle Report	1 day?	28/07/10 08:(
PROJECT WRITE UP	13 days?	29/07/10 08:(
Supervisor Away from University	11 days?	03/08/10 08:(
Student Away	2 days?	06/08/10 08:(
Report Contingency Time!!!	10 days?	16/08/10 08:(
Supervisor Meeting - Final Comments	1 day?	24/08/10 08:(
DEADLINE: PROJECT COMPLETION!!!!!	1 day?	27/08/10 08:(

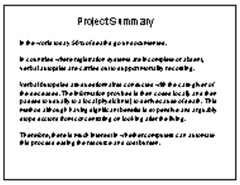




Appendix I: Presentation Delivered at Progress Meeting, July 2010.

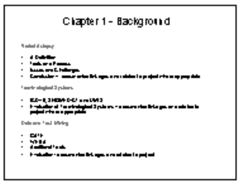


1

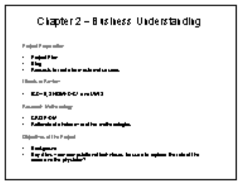


2

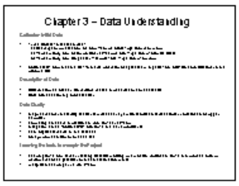
3



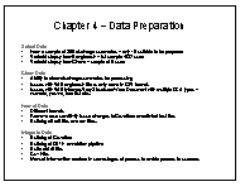
4



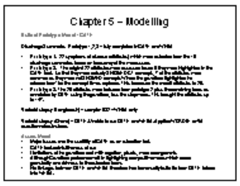
5



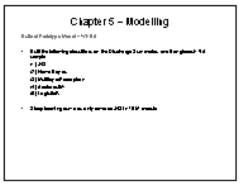
6



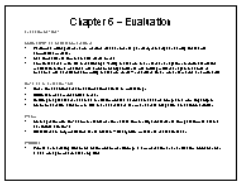
7



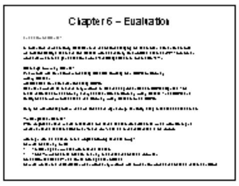
8



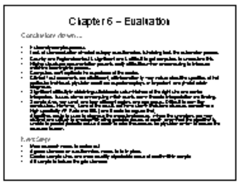
9



10



11



12

Appendix J: National Institute of Health (NIH) Certificate "Protecting Human Research Participants

Certificate of Completion

The National Institutes of Health (NIH) Office of Extramural Research certifies that **rebecca west** successfully completed the NIH Web-based training course "Protecting Human Research Participants".

Date of completion: 05/08/2010

Certification Number: 444708

Appendix K: Sample Discharge Summary

Discharged :▯**DATE[Sep 29 2007]▯Dict :▯**NAME[XXX , WWW]▯Attend :▯**NAME[ZZZ , YYY]▯PRINCIPAL DIAGNOSES :▯1. Exacerbation of
congestive heart failure .▯2. Exacerbation of chronic obstructive pulmonary disease .▯SECONDARY DIAGNOSES :▯1. Hypertension
.▯2. Status post myocardial infarction .▯3. Status post transient ischemic attack .▯4. Status post deep venous thrombosis .▯
5. History of peripheral vascular disease .▯6. Arthritis .▯7. History of renal cell carcinoma status post left nephrectomy
with chronic renal insufficiency .▯8. Cholecystitis .▯9. Status post carotid endarterectomy .▯10. Status post fem-pop bypass
bilaterally x 3 .▯11. Status post abdominal aortic aneurysm repair .▯LIST OF DISCHARGE MEDICATIONS :▯1. Prednisone 10 mg 1
tablet p.o. b.i.d. x 3 days 1 tablet p.o. daily 3 days .▯2. Coumadin 2 mg 1 tablet p.o. daily at night .▯3. Spiriva 1 puff
daily .▯4. Labetalol 300 mg 1 tablet p.o. b.i.d.▯5. Isosorbide mononitrate TR 30 mg 1 tablet p.o. daily .▯6. Lasix 40 mg 1
tablet p.o. daily .▯7. Advair Diskus 250/50 1 puff b.i.d.▯8. Doxycycline 100 mg 1 tablet p.o. b.i.d. x 7 days .▯9. Enteric
coated aspirin 81 mg 1 tablet p.o. daily .▯10. Norvasc 10 mg 1 tablet daily .▯11. INR / PT / BMP weekly via home care ,
oxygen at home care .▯PRINCIPAL TESTS AND PROCEDURES :▯Lower extremity Dopplers negative for deep venous thrombosis in either
leg bilaterally .▯LIST OF CONSULTANTS :▯1. Cardiology .▯2. Home Care .▯HOSPITAL COURSE AND TREATMENT :▯Briefly , this is a
**AGE[in 70s]- year - old female with a past medical history significant for congestive heart failure , COPD , and history of
DVT , who presented with right calf swelling and increased shortness of breath x 3 days .▯1. Shortness of breath :▯We felt
that the patient 's shortness of breath was either likely due to CHF and / or COPD exacerbation .▯Her PE was unlikely due to
the fact the patient was on Coumadin as an outpatient , and had been anticoagulated for some amount of time .▯This was
reviewed with the attending physician , Dr. **NAME[YYY ZZZ] , who agreed that a spiral CT should not be performed .▯The
patient was admitted to the general medicine floor .▯Her BNP was checked and found to be 347 .▯The patient was treated with
p.o. Lasix and was diuresed .▯She was continued on her outpatient medications Spiriva and Advair as well as given DuoNeb as
needed for any shortness of breath .▯The patient was also treated for presumptively for her COPD exacerbation , was started
on a 10 - day course of doxycycline as well as started on oral prednisone .▯A Cardiology consult was also obtained to rule
out any cardiac causes of the patient 's shortness of breath .▯Per Cardiology , it was felt that the patient had no clinical
evidence of any CHF by physical examination or x-ray ; however , the patient did improve subjectively on Lasix , prednisone
as well as doxycycline .▯It was recommend to continue current treatment .▯It was felt that the patient 's mildly elevated BMP
was secondary to pulmonary hypertension seen on the patient 's last echocardiogram at last admission .▯After treatment for
CHF and COPD exacerbation , the patient 's shortness of breath improved .▯She was at her baseline shortness of breath at home
.▯2. Cardiology .▯The patient was continued on her outpatient antihypertensive medications throughout her hospital course .▯
The patient 's blood pressure was within normal limits , and she was hemodynamically stable .▯3. Renal .▯The patient does
have a history of renal insufficiency status post left nephrectomy due to renal cell carcinoma .▯It was felt that she could
be followed up on an outpatient basis for her renal insufficiency , and her creatinine was at her baseline in 2006 , at 1.5
.▯4. Prophylaxis .▯The patient was treated with Pepcid and heparin for GIT and DVT prophylaxis .▯During the course of the
hospital stay , the patient was afebrile .▯Her vital signs were stable .▯She had no significant complaints .▯At the time of
discharge , the patient did not complain of any significantly new shortness of breath .▯She stated that she was feeling
better .▯DISPOSITION :▯The patient is to be discharged to home on the medications as listed above .▯She is instructed to
resume activity as tolerated and also to resume a cardiac healthy diet .▯Home Care consults were obtained for the patient to
have PT / INR / BMP checks weekly as well as to ensure the patient had an appropriate home O2 therapy .▯At the time of
discharge , the patient was afebrile .▯Her vital signs were stable , and she had no significant complaints .▯The patient
agreed that she was at her baseline .▯The patient was ambulating appropriately and tolerating a p.o. diet .▯The patient was
instructed to contact her primary care physician and / or contact the emergency room if she had any symptoms including but
not limited to prolonged fever , nausea , vomiting , chills , night sweats , chest pain , shortness of breath , palpitations
or any other serious complaints .▯______________________________▯**NAME[WWW XXX]▯HS Job # 123616 / 41427 / 40215▯SHY #
ID-NUM▯D :▯DATE[Sep 29 2007] 09:22▯T :▯**DATE[Sep 30 2007] 11:26▯**CARBON-COPY▯

Appendix L: Gold Standards for Cause of Death Diagnoses:Ghana Verbal Autopsies

No	Infanid	Code	Cause
1	KKL0254/22C1	22	Severe infection
2	KAJ0149/04C1	21	Prematurity
3	KJM0202/29C1	29	Unexplained
4	KD047/2/33C1	22	Severe infection
5	KA061/2/08C1	26	Congenital anomaly

Appendix M: Extract from SNOMED Concept File

```
CONCEPTID       CONCEPTSTATUS   FULLYSPECIFIEDNAME      CTV3ID  SNOMEDID        ISPRIMITIVE
139784008       0       Entire tuberculum sellae (body structure)       XS10s   T-D1463 1
100419000       10      DUOVAC -M (product)     XU07K   C-D2631 1
140087001       0       Entire clivus ossis sphenoidalis (body structure)       XS1BZ   T-11183 1
100331002       10      DERMCAPS ES LIQUID (product)    XU05n   C-D2411 1
100334005       10      DERMOLAR SHAMPOO (product)      XU05q   C-D2417 1
100361005       10      DIFIL SYRUP (product)   XU06K   C-D2499 1
100362003       10      DIFIL TABS (product)    XU06L   C-D2501 1
100390004       10      DL-ALPHA TOCOPHEROL ACETATE INJECTION (product) XU06p   C-D2569 1
10039002        0       ^210m^Bismuth (substance)       XU06q   C-125B2 1
100391000       10      D-LIMONENE SHAMPOO (product)    XU06r   C-D2571 1
100420006       10      DUOVAC -Ma5 (product)   XU07L   C-D2633 1
10042008        0       Structure of intervertebral foramen of fifth thoracic vertebra (body structure) XU07M   T-1175A 1
100335006       10      DEXAMETHASONE 2.0 MG INJECTION (product)        XU05r   C-D2421 1
100336007       10      DEXAMETHASONE INJECTION (product)       XU05s   C-D2423 1
100363008       10      DINEOTEX (product)      XU06M   C-D2503 1
100364002       10      DIOCTYNATE (product)    XU06N   C-D2505 1
100392007       10      D-L-M TABLETS (product) XU06s   C-D2573 1
100393002       10      d-L-METHIONINE POWDER (product) XU06t   C-D2575 1
140895003       0       Entire sphenooccipital synchondrosis (body structure)   XS03L   T-15261 1
100500008       10      ENTERITIS FORMULA (product)     XU08k   C-D2827 1
100526003       10      EQUINIME (product)      XU09C   C-D2887 1
100527007       10      EQUIPAR EQUINE WORMER PASTE (product)   XU09D   C-D2889 1
100449003       10      DYNATABS (product)      XU07r   C-D2701 1
140390001       0       Entire pterygoid process of sphenoid bone (body structure)      XS0Js   T-1119B 1
100476003       10      EFA LIQUID (product)    XU08L   C-D2767 1
100477007       10      EFA-Z PLUS (product)    XU08M   C-D2769 1
10050004        0       Contusion of chest (disorder)   SE21.   DD-53310        0
1005009 0       Entire diaphragmatic lymph node (body structure)        XS0WA   T-C4380 1
100528002       10      EQUIPAR EQUINE WORMER SUSPENSION (product)      XU09E   C-D2891 1
100529005       10      EQUI-PHAR DL-METHIONINE POWDER (product)        XU09F   C-D2893 1
100450003       10      DYNATABS -T (product)   XU07s   C-D2705 1
100451004       10      DYNE (product)  XU07t   C-D2707 1
100478002       10      ELECTROID 7 (product)   XU08N   C-D2771 1
100479005       10      ELECTROID 7 PLUS H.S. (product) XU08O   C-D2773 1
100501007       10      ENTERO-GUARD (product)  XU08l   C-D2829 1
100502000       10      ENTROLYTE (product)     XU08m   C-D2831 1
```

Appendix N: Requests and Questions on IHME Verbal Autopsy Data

mscgirl
July 7, 2010 at 12:15 pm
Hi Abraham, I wonder if you can help me. I'm an MSc Student @ Leeds University undertaking my project in tagging medical concepts with verbal autopsies. Please see my blog http://mscgirl.wordpress.com/. I was looking for anonymised verbal autopsy data and found some on your site, which was v.interesting. I took the data and loaded it into a machine learning tool (WEKA) but unfortunately it didn't mean too much to me as I could not ascertain what the symptoms and cause of death were, as they were numeric. Is it possible that you could advise me of this information? It would very much help me with my project as with your permission I would like to use your data to explain my technique of concept extraction and classification that I have developed. Thanking you in advance, Rebecca
Abraham Flaxman
July 7, 2010 at 5:18 pm
Hi Rebecca,
I'm working on getting the full data released publicly for people like you to use. But it might take a while…I'll move this conversation to the verbal autopsy challenge page, but I wanted to reply here to make sure you got it.
Rebecca
July 7, 2010 at 6:46 pm
Thanks Abraham for your quick response. My project needs to conclude by end of August of this year. So I guess it will not be fully available by then for me to use. Although you never know. In the meantime I will try and pursue some other avenues. Thanks for your post and your article I found it a real interesting read
Rebecca
July 27, 2010 at 10:57 am
Hi Abraham, still working on my verbal autopsy project and I'm still looking to use the dataset that you used in your paper. I know you have explained that you cannot give out all the details on the symptoms and I do fully understand and accept this. However, to allow me to interpret the csv file it would be very helpful for me to understand which columns are the actual symptoms of the diseases. In my project I am trying to take the disease symptoms and then run them through various classifiers to see how accurate it predicts probable cause of death. At present when I upload the file into WEKA I am getting some very strange results. In your paper you say the file has 928 rows, 1528 attributes of which 200 actual correspond to VA survey questions and they are 140 causes of death. So to help me please could you advise which columns are disease symptoms it would help me enormously to make sense of the data. Finally in your paper you say that there 140 cause of death. In column "EM" annotated "cause of death" there are numbers 1-32 so I interpreted this that there were 32 causes of death that were categorized. Please could you explain, I must be missing something? I apologise for all the questions, but this is the first sample that I have come across that looks very promising indeed and is of a suitable size. I have been many places to get VA data and have struggled enormously. The best that I have been able to get is 5 VA's from Ghana. So as you can see I have a real problem with sample size! Thank you for reading and hoping you can help. Rebecca
Sean
July 27, 2010 at 6:05 pm
Hi Rebecca,
I worked on the verbal autopsy paper with Abie and I think I can answer some of your questions. The symptoms are a mixture of categorical, continuous, and binary data. If it helps I can let you know the following:
1) symptom2 is an age variable and should be treated as continuous
2) symptoms 27, 40, 45, 73, 77, 81, 83, 90, and 138 all describe the duration of symptoms listed elsewhere in the survey and should also be treated as continuous.
3) symptom 140 is a location variable and should be treated as categorical.
4) All other symptoms should be treated as categorical. If the symptom values are binary, then it is a yes/no question. If the values are integers and include several different values then it is a symptom question with many categories.
5)For any of the symptoms there are two special values you should take note of:
a) A value of "99″ indicates "did not know"
b) A value of "-1″ indicates "no response"
In the paper we state that the sample Bangladesh data set at measureddhs.com has 928 rows, 1528 attributes, and 140 causes of death; however, the Bangladesh data set is not the one we posted. The data we posted contains anonymized data from another country. It has only 142 symptom questions (if you consider age, location, and duration to be symptoms) and has only 32 unique causes of death.
So you were correct when you determined that there are 32 causes of death.
I hope this helps!

Correspondence with Abraham Flaxman and Sean Green authors of
Machine Learning Methods for Verbal Autopsy in Developing Countries.
(2009). *Association for the Advancement of Artificial Intelligence.*

Appendix O: Python Program to change case of SNOMED-CT concept file.

```
f=open('concept-results.txt','r')
f2 = open('aaaa.txt','w')
deslist = []
for line in f.readlines():
    newline = line.lower()
    deslist.append(newline)

f2.writelines(deslist)

f.close()
f2.close()
```

Appendix P: ARFF file Example

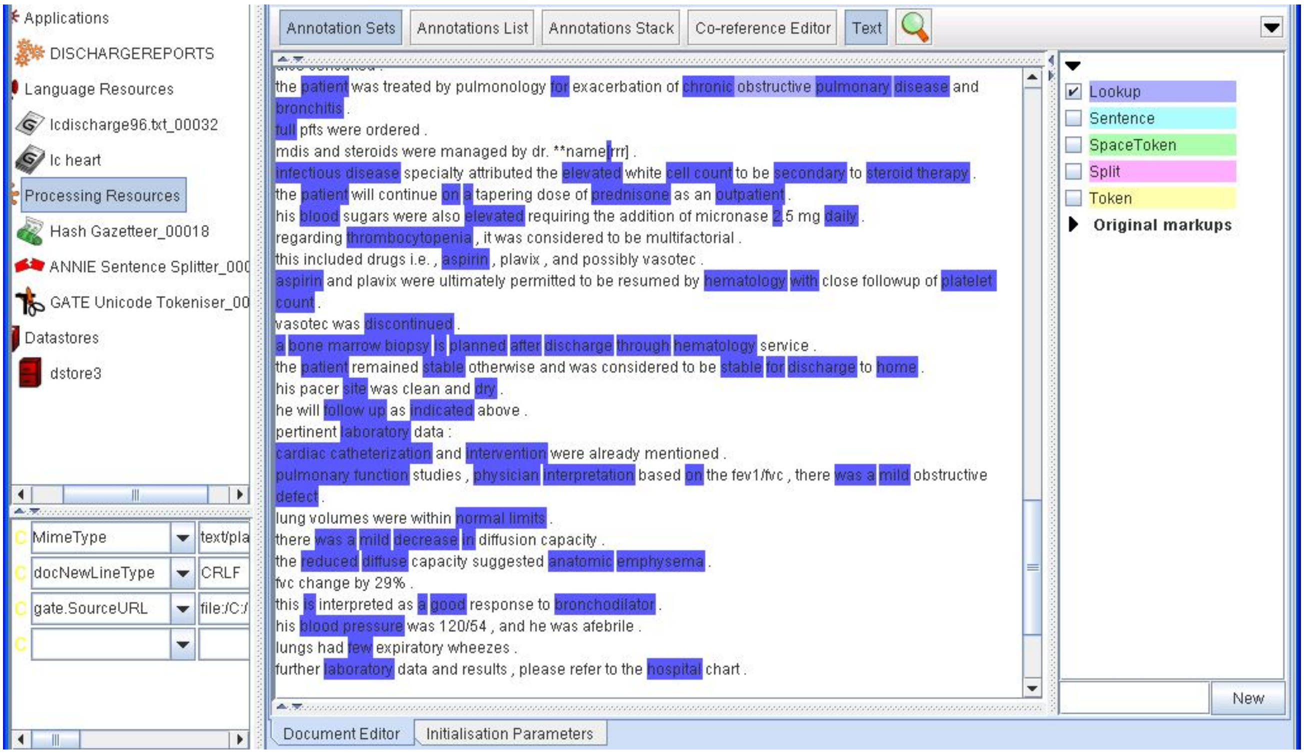
```
% 1. Title: VA Database
%
% 2. Sources:
%      (a) Creator: R.L West
%      (b) Donor: i2b2 Challenge.
%      (c) Date: July, 2010
%
%
% 3. Number of Instances: 16 (three classes)
%
% 4. Number of Attributes: 21 real, predictive attributes and the class
%
%
% 5. Missing Attribute Values: None

@RELATION va

@ATTRIBUTE cough                REAL
@ATTRIBUTE coughing             REAL
@ATTRIBUTE pleural              REAL
@ATTRIBUTE effusion             REAL
@ATTRIBUTE lobe                 REAL
@ATTRIBUTE sputum               REAL
@ATTRIBUTE fluid                REAL
@ATTRIBUTE WBC                  REAL
@ATTRIBUTE lung                 REAL
@ATTRIBUTE chest                REAL
@ATTRIBUTE angina               REAL
@ATTRIBUTE shortnessofbreath    REAL
@ATTRIBUTE hypertension         REAL
@ATTRIBUTE infarction           REAL
@ATTRIBUTE blood                REAL
@ATTRIBUTE pressure             REAL
@ATTRIBUTE artery               REAL
@ATTRIBUTE catheterization      REAL
@ATTRIBUTE chestxray            REAL
@ATTRIBUTE fever                REAL
@ATTRIBUTE chills               REAL

@ATTRIBUTE class        {Pneumonia,CoronaryArteryDisease,ChronicObstructivePulmonaryDisease}

@DATA
0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,2,0,1,0,Pneumonia
1,0,0,2,2,0,2,0,2,2,3,0,0,0,0,1,1,0,0,2,1,Pneumonia
0,0,0,0,1,0,0,0,1,0,0,0,2,0,2,1,0,0,1,1,0,Pneumonia
0,0,0,0,0,0,0,0,1,0,0,0,1,0,0,0,0,0,1,0,0,Pneumonia
0,0,0,0,1,2,0,0,1,0,0,1,0,0,0,0,0,0,3,0,0,Pneumonia
2,1,1,2,1,0,0,0,1,1,0,2,0,0,0,0,0,0,2,0,0,Pneumonia
0,0,2,2,0,0,2,0,2,3,0,0,0,0,1,1,1,0,4,0,0,Pneumonia
0,1,4,6,1,2,8,2,0,3,0,0,0,0,1,0,1,0,6,0,0,Pneumonia
0,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,ChronicObstructivePulmonaryDisease
0,0,0,0,0,0,0,0,0,0,0,8,1,1,1,1,1,0,0,0,0,ChronicObstructivePulmonaryDisease
0,0,1,2,0,1,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,ChronicObstructivePulmonaryDisease
0,0,0,0,0,0,0,0,0,0,0,1,1,0,1,0,0,0,0,0,0,CoronaryArteryDisease
0,0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,1,0,0,0,CoronaryArteryDisease
0,0,0,0,0,0,0,0,0,0,0,0,1,2,1,0,1,5,0,0,0,CoronaryArteryDisease
0,0,0,0,0,0,0,0,0,1,0,0,2,0,1,1,1,1,0,0,0,CoronaryArteryDisease
0,0,0,0,0,0,0,0,0,0,1,0,1,0,0,0,4,0,0,0,0,CoronaryArteryDisease

%
```

Appendix Q: Example of an Annotated Discharge Summary in GATE

Un-annotated:

```
Discharged :▯**DATE[Sep 29 2007]▯Dict :▯**NAME[XXX , WWW]▯Attend :▯**NAME[ZZZ , YYY]▯PRINCIPAL DIAGNOSES :▯1. Exacerbation of
congestive heart failure .▯2. Exacerbation of chronic obstructive pulmonary disease .▯SECONDARY DIAGNOSES :▯1. Hypertension
.▯2. Status post myocardial infarction .▯3. Status post transient ischemic attack .▯4. Status post deep venous thrombosis .▯
5. History of peripheral vascular disease .▯6. Arthritis .▯7. History of renal cell carcinoma status post left nephrectomy
with chronic renal insufficiency .▯8. Cholecystitis .▯9. Status post carotid endarterectomy .▯10. Status post fem-pop bypass
bilaterally x 3 .▯11. Status post abdominal aortic aneurysm repair .▯LIST OF DISCHARGE MEDICATIONS :▯1. Prednisone 10 mg 1
tablet p.o. b.i.d. x 3 days 1 tablet p.o. daily 3 days .▯2. Coumadin 2 mg 1 tablet p.o. daily at night .▯3. Spiriva 1 puff
daily .▯4. Labetalol 300 mg 1 tablet p.o. b.i.d.▯5. Isosorbide mononitrate TR 30 mg 1 tablet p.o. daily .▯6. Lasix 40 mg 1
tablet p.o. daily .▯7. Advair Diskus 250/50 1 puff b.i.d.▯8. Doxycycline 100 mg 1 tablet p.o. b.i.d. x 7 days .▯9. Enteric
coated aspirin 81 mg 1 tablet p.o. daily .▯10. Norvasc 10 mg 1 tablet daily .▯11. INR / PT / BMP weekly via home care ,
oxygen at home care .▯PRINCIPAL TESTS AND PROCEDURES :▯Lower extremity Dopplers negative for deep venous thrombosis in either
leg bilaterally .▯LIST OF CONSULTANTS :▯1. Cardiology .▯2. Home Care .▯HOSPITAL COURSE AND TREATMENT :▯Briefly , this is a
**AGE[in 70s]- year - old female with a past medical history significant for congestive heart failure , COPD , and history of
DVT , who presented with right calf swelling and increased shortness of breath x 3 days .▯1. Shortness of breath :▯We felt
that the patient 's shortness of breath was either likely due to CHF and / or COPD exacerbation .▯Her PE was unlikely due to
the fact the patient was on Coumadin as an outpatient , and had been anticoagulated for some amount of time .▯This was
reviewed with the attending physician , Dr. **NAME[YYY ZZZ] , who agreed that a spiral CT should not be performed .▯The
patient was admitted to the general medicine floor .▯Her BNP was checked and found to be 347 .▯The patient was treated with
p.o. Lasix and was diuresed .▯She was continued on her outpatient medications Spiriva and Advair as well as given DuoNeb as
needed for any shortness of breath .▯The patient was also treated for presumptively for her COPD exacerbation , was started
on a 10 - day course of doxycycline as well as started on oral prednisone .▯A Cardiology consult was also obtained to rule
out any cardiac causes of the patient 's shortness of breath .▯Per Cardiology , it was felt that the patient had no clinical
evidence of any CHF by physical examination or x-ray ; however , the patient did improve subjectively on Lasix , prednisone
as well as doxycycline .▯It was recommend to continue current treatment .▯It was felt that the patient 's mildly elevated BMP
was secondary to pulmonary hypertension seen on the patient 's last echocardiogram at last admission .▯After treatment for
CHF and COPD exacerbation , the patient 's shortness of breath improved .▯She was at her baseline shortness of breath at home
.▯2. Cardiology .▯The patient was continued on her outpatient antihypertensive medications throughout her hospital course .▯
The patient 's blood pressure was within normal limits , and she was hemodynamically stable .▯3. Renal .▯The patient does
have a history of renal insufficiency status post left nephrectomy due to renal cell carcinoma .▯It was felt that she could
be followed up on an outpatient basis for her renal insufficiency , and her creatinine was at her baseline in 2006 , at 1.5
.▯4. Prophylaxis .▯The patient was treated with Pepcid and heparin for GIT and DVT prophylaxis .▯During the course of the
hospital stay , the patient was afebrile .▯Her vital signs were stable .▯She had no significant complaints .▯At the time of
discharge , the patient did not complain of any significantly new shortness of breath .▯She stated that she was feeling
better .▯DISPOSITION :▯The patient is to be discharged to home on the medications as listed above .▯She is instructed to
resume activity as tolerated and also to resume a cardiac healthy diet .▯Home Care consults were obtained for the patient to
have PT / INR / BMP checks weekly as well as to ensure the patient had an appropriate home O2 therapy .▯At the time of
discharge , the patient was afebrile .▯Her vital signs were stable , and she had no significant complaints .▯The patient
agreed that she was at her baseline .▯The patient was ambulating appropriately and tolerating a p.o. diet .▯The patient was
instructed to contact her primary care physician and / or contact the emergency room if she had any symptoms including but
not limited to prolonged fever , nausea , vomiting , chills , night sweats , chest pain , shortness of breath , palpitations
or any other serious complaints .▯______________________________▯**NAME[WWW XXX]▯HS Job # 123616 / 41427 / 40215▯SHY #
**ID-NUM▯D :▯**DATE[Sep 29 2007] 09:22▯T :▯**DATE[Sep 30 2007] 11:26▯**CARBON-COPY▯
```

Annotated:



Appendix R: Python Code to read and extract SNOMED-CT Concepts

Discharge Summaries:

```
import glob
import os
wd = os.getcwd ()
foldername = "dischargesummary"
path = wd + "\\" + foldername
os.chdir(path)

#print path

files = glob.glob('*.txt')
for fle in files:
    print '\n\n', fle

    f = open(fle, "r")

    termdict = {'cough':0, 'coughing':0, 'pulmonary hypertension':0,
    'effusion':0,'blood pressure':0, 'sputum':0, 'fluid':0, 'wbc':0,
    'pleural effusion':0, 'myocardinal infarction':0, 'angina':0,
    'shortness of breath':0, 'white sputum':0,'cardiac catherization':0,
    'blood':0, 'coronary artery':0, 'catheterization':0, 'chest xray':0,
    'fever':0, 'white sputum':0, 'renal artery stenosis':0,
    'pericardial effusion':0, 'green sputum':0, 'chill':0}

    termlist = ["cough", "coughing", "pulmonary hypertension",
    "effusion", "blood pressure", "sputum", "fluid", "wbc","pleural effusion",
    "myocardinal infarction", "angina", "shortness of breath", "white sputum",
    "cardiac catherization", "blood", "coronary artery", "catheterization",
    "chest xray","fever", "white sputum", "renal artery stenosis",
    "pericardial effusion","green sputum", "chill"]


    lines = f.readlines()
    for line in lines:
        for term in termlist:
            newline = line.lower()
            occurslst = newline.split(term)
            occursint = len(occurslst) - 1
            termdict[term] = termdict[term] + occursint



    for key in termdict:
        print key, termdict[key],

    print '\n'


    for key in termdict:
        print (key,),

    print '\n'

    for key in termdict:
        print str(termdict[key]) + ",",


    f.close()
```

Appendix R Continued … Verbal Autopsies:

```
import glob
import os
wd = os.getcwd ()
foldername = "kintampo"
path = wd + "\\" + foldername
os.chdir(path)

#print path

files = glob.glob('*.txt')
for fle in files:
    print '\n\n', fle

    f = open(fle, "r")

    termdict = {'malaria':0, 'severe headache':0, 'paracetamol':0,
    'chloroquine':0, 'lower abdomen':0, 'fluid':0,'convulsion':0,
    'herbal medicine':0,'weak':0, 'antenatal':0, 'water':0, 'bleeding':0,
    'medical assistant':0, 'traditional':0,'started crying':0, 'crying':0,
    'jaw':0, 'suck':0, 'blood test':0, 'hospital':0, 'child':0,
    'bottom in two':0, 'health centre':0, 'dead':0, 'severe headache':0,
    'hospital':0, 'dizziness':0, 'illness':0, 'blood':0, 'birth':0,
    'water':0, 'started breastfeeding':0, 'injection':0, 'fever':0,


    'antenatal':0, 'pain':0, 'hospital illness':0, 'able to suck':0,
    'accident':0, 'bleeding':0, 'clinic':0, 'medical assistant':0,
    'traditional':0, 'pregnancy':0, 'normal delivery':0, 'condition':0,
    'sickness':0, 'death':0, 'bulging':0, 'fontanel':0, 'sucking':0,}

    termlist = ["malaria", "severe headache", "paracetamol", "chloroquine",
    "lower abdomen", "fluid", "convulsion", "herbal medicine", "weak",
    "antenatal", "water", "bleeding", "medical assistant", "traditional",
    "started crying", "crying", "jaw", "suck", "blood test", "hospital",
    "child", "bottom in two", "health centre", "dead", "severe headache"
    "hospital", "dizziness", "illness", "blood", "birth", "water",
    "started breastfeeding" "injection", "fever", "antenatal", "pain",
    "hospital illness", "able to suck", "accident", "bleeding", "clinic",
    "medical assistant", "traditional", "pregnancy", "normal delivery",
    "condition","sickness", "death", "bulging", "fontanel", "sucking"]


    lines = f.readlines()
    for line in lines:
        for term in termlist:
            newline = line.lower()
            occurslst = newline.split(term)
            occursint = len(occurslst) - 1
```

```
            newline = line.lower()
            occurslst = newline.split(term)
            occursint = len(occurslst) - 1
            termdict[term] = termdict[term] + occursint



    for key in termdict:
        print key, termdict[key],

    print '\n'

    for key in termdict:
        print (key,),

    print '\n'

    for key in termdict:
        print str(termdict[key]) + ",",


    f.close()
```

APPENDIX S: WEKA Results for US Discharge Summaries

Measurement - OneR Cross Validation	Prototype 1			Prototype 2			Prototype 3		
Number of Attributes	21			24			146		
Total Number of Instances	16			16			16		
No: Correctly Classified Instances	10			12			12		
No: Incorrectly Classified Instances	6			4			4		
% Correctly Classified Instances	62.5%			75%			75%		
% Incorrectly Classified Instances	37.5%			25%			25%		
TP Rate Pneumonia	0.75			1			1		
TP Rate Coronary Artery Disease	0.8			0.2			0.2		
TP Rate Chronic Obstruction Pulmonary Disease	0			1			1		
FP Rate Pneumonia	0.375			0.5			0.5		
FP Rate Coronary Artery Disease	0.273			0			0		
FP Rate Chronic Obstruction Pulmonary Disease	0			0			0		
Precision Pneumonia	0.667			0.667			0.667		
Precision Coronary Artery Disease	0.571			1			1		
Precision Chronic Obstruction Pulmonary Disease	0			1			1		
Recall Pneumonia	0.75			1			1		
Recall Coronary Artery Disease	0.8			0.2			0.2		
Recall Chronic Obstruction Pulmonary Disease	0			1			1		
Confusion Matrix	**Prototype 1**			**Prototype 2**			**Prototype 3**		
Classified As	a	b	c	a	b	c	a	b	c
Chronic Obstructive Pulmonary Disease = a	3	0	0	3	0	0	3	0	0
Coronary Artery Disease = b	0	1	4	0	1	4	0	1	4
Pneumonia = c	0	0	8	0	0	8	0	0	8

Measurement - ZeroR Cross Validation	**Prototype 1**			**Prototype 2**			**Prototype 3**		
Number of Attributes	21			24			146		
Total Number of Instances	16			16			16		
No: Correctly Classified Instances	8			8			8		
No: Incorrectly Classified Instances	8			8			8		
% Correctly Classified Instances	50%			50%			50%		
% Incorrectly Classified Instances	50%			50%			50%		
TP Rate Pneumonia	1			1			1		
TP Rate Coronary Artery Disease	0			0			0		
TP Rate Chronic Obstruction Pulmonary Disease	0			0			0		
FP Rate Pneumonia	1			1			1		
FP Rate Coronary Artery Disease	0			0			0		
FP Rate Chronic Obstruction Pulmonary Disease	0			0			0		
Precision Pneumonia	0.5			0.5			0.5		
Precision Coronary Artery Disease	0			0			0		
Precision Chronic Obstruction Pulmonary Disease	0			0			0		
Recall Pneumonia	1			1			1		
Recall Coronary Artery Disease	0			0			0		
Recall Chronic Obstruction Pulmonary Disease	0			0			0		
Confusion Matrix	**Prototype 1**			**Prototype 2**			**Prototype 3**		
Classified As	a	b	c	a	b	c	a	b	c
Chronic Obstructive Pulmonary Disease = a	0	0	3	0	0	3	0	0	3
Coronary Artery Disease = b	0	0	5	0	0	5	0	0	5
Pneumonia = c	0	0	8	0	0	8	0	0	8

Measurement - J-Rip Cross Validation	Prototype 1	Prototype 2	Prototype 3
Number of Attributes	21	24	146
Total Number of Instances	16	16	16
No: Correctly Classified Instances	7	10	11
No: Incorrectly Classified Instances	9	6	5
% Correctly Classified Instances	43.75%	62.5%	68.75%
% Incorrectly Classified Instances	56.25%	37.5%	31.25%
TP Rate Pneumonia	0.75	0.75	1
TP Rate Coronary Artery Disease	0.2	0.8	0.6
TP Rate Chronic Obstruction Pulmonary Disease	0	0	0
FP Rate Pneumonia	0.5	0.125	0.375
FP Rate Coronary Artery Disease	0.4	0.455	0.091
FP Rate Chronic Obstruction Pulmonary Disease	0	0	0.077
Precision Pneumonia	0.6	0.857	0.727
Precision Coronary Artery Disease	0.167	0.444	0.75
Precision Chronic Obstruction Pulmonary Disease	0	0	0
Recall Pneumonia	0.75	0.5	1
Recall Coronary Artery Disease	0.2	0.8	0.6
Recall Chronic Obstruction Pulmonary Disease	0	0	0

Confusion Matrix	Prototype 1			Prototype 2			Prototype 3		
Classified As	a	b	c	a	b	c	a	b	c
Chronic Obstructive Pulmonary Disease = a	0	3	0	0	3	0	0	1	2
Coronary Artery Disease = b	0	1	4	0	4	1	1	3	1
Pneumonia = c	0	2	6	0	2	6	0	0	8

Measurement - J48 Cross Validation	**Prototype 1**	**Prototype 2**	**Prototype 3**
Number of Attributes	21	24	146
Total Number of Instances	16	16	16
No: Correctly Classified Instances	12	12	5
No: Incorrectly Classified Instances	4	4	5
% Correctly Classified Instances	75%	75%	68.75%
% Incorrectly Classified Instances	25%	25%	31.25%
TP Rate Pneumonia	1	1	1
TP Rate Coronary Artery Disease	0.8	0.8	0.6
TP Rate Chronic Obstructive Pulmonary Disease	0	0	0
FP Rate Pneumonia	0	0	0
FP Rate Coronary Artery Disease	0.273	0.273	0.273
FP Rate Chronic Obstructive Pulmonary Disease	0.077	0.077	0.154
Precision Pneumonia	1	1	1
Precision Coronary Artery Disease	0.571	0.571	0.5
Precision Chronic Obstructive Pulmonary Disease	0	0	0
Recall Pneumonia	1	1	1
Recall Coronary Artery Disease	0.667	0.667	0.6
Recall Chronic Obstructive Pulmonary Disease	0	0	0

Confusion Matrix	**Prototype 1**			**Prototype 2**			**Prototype 3**		
Classified As	a	b	c	a	b	c	a	b	c
Chronic Obstructive Pulmonary Disease = a	0	3	0	0	3	0	0	3	0
Coronary Artery Disease = b	1	4	0	1	4	0	2	3	0
Pneumonia = c	0	0	8	0	0	8	0	0	8

Measurement- Naïve Bayes (Cross-Validation)	Prototype 1			Prototype 2			Prototype 3		
Number of Attributes	21			24			146		
Total Number of Instances	16			16			16		
No: Correctly Classified Instances	10			11			7		
No: Incorrectly Classified Instances	6			5			9		
% Correctly Classified Instances	62.5%			68.75%			43.75%		
% Incorrectly Classified Instances	37.5%			31.25%			56.25%		
TP Rate Pneumonia	0.875			1			0.875		
TP Rate Coronary Artery Disease	0.6			0.6			0		
TP Rate Chronic Obstructive Pulmonary Disease	0			0			0		
FP Rate Pneumonia	0.375			0.125			0.875		
FP Rate Coronary Artery Disease	0.182			0.182			0.182		
FP Rate Chronic Obstructive Pulmonary Disease	0.077			0.154			0		
Precision Pneumonia	0.7			0.889			0.5		
Precision Coronary Artery Disease	0.6			0.6			0		
Precision Chronic Obstructive Pulmonary Disease	0			0			0		
Recall Pneumonia	0.875			1			0.875		
Recall Coronary Artery Disease	0.6			0.6			0		
Recall Chronic Obstructive Pulmonary Disease	0			0			0		
Confusion Matrix	**Prototype 1**			**Prototype 2**			**Prototype 3**		
Classified As	a	b	c	a	b	c	a	b	c
Chronic Obstructive Pulmonary Disease = a	0	1	2	0	2	1	0	1	2
Coronary Artery Disease = b	1	3	1	2	3	0	0	0	5
Pneumonia = c	0	1	7	0	0	8	0	1	7

Measurement - MultiLayerPerceptron (Cross-Val)	Prototype 1	Prototype 2	Prototype 3
Number of Attributes	21	24	146
Total Number of Instances	16	16	16
No: Correctly Classified Instances	8	8	8
No: Incorrectly Classified Instances	8	8	8
% Correctly Classified Instances	50%	50%	50%
% Incorrectly Classified Instances	50%	50%	50%
TP Rate Pneumonia	0.625	0.875	0.875
TP Rate Coronary Artery Disease	0.6	0.2	0.2
TP Rate Chronic Obstructive Pulmonary Disease	0	0	0
FP Rate Pneumonia	0.25	0.5	0.875
FP Rate Coronary Artery Disease	0.273	0.273	0.091
FP Rate Chronic Obstructive Pulmonary Disease	0.231	0.077	0
Precision Pneumonia	0.714	0.636	0.5
Precision Coronary Artery Disease	0.5	0.25	0.5
Precision Chronic Obstructive Pulmonary Disease	0	0	0
Recall Pneumonia	0.625	0.875	0.875
Recall Coronary Artery Disease	0.6	0.2	0.2
Recall Chronic Obstructive Pulmonary Disease	0	0	0

Confusion Matrix	Prototype 1			Prototype 2			Prototype 3		
Classified As	a	b	c	a	b	c	a	b	c
Chronic Obstructive Pulmonary Disease = a	0	1	2	0	2	1	0	0	3
Coronary Artery Disease = b	2	3	0	1	1	3	0	1	4
Pneumonia = c	1	2	5	0	1	7	0	1	7

Measurement -AdaboostM1 (Cross-Val)	Prototype 1	Prototype 2	Prototype 3
Number of Attributes	21	24	146
Total Number of Instances	16	16	16
No: Correctly Classified Instances	12	16	16
No: Incorrectly Classified Instances	4	0	0
% Correctly Classified Instances	75.0%	100%	100%
% Incorrectly Classified Instances	25.0%	0%	0%
TP Rate Pneumonia	1	1	1
TP Rate Coronary Artery Disease	0.8	1	1
TP Rate Chronic Obstructive Pulmonary Disease	0	1	1
FP Rate Pneumonia	0	0	0
FP Rate Coronary Artery Disease	0.273	0	0
FP Rate Chronic Obstructive Pulmonary Disease	0.077	0	0
Precision Pneumonia	1	1	1
Precision Coronary Artery Disease	0.571	1	1
Precision Chronic Obstructive Pulmonary Disease	0	1	1
Recall Pneumonia	1	1	1
Recall Coronary Artery Disease	0.8	1	1
Recall Chronic Obstructive Pulmonary Disease	0	1	1

Confusion Matrix	Prototype 1			Prototype 2			Prototype 3		
Classified As	a	b	c	a	b	c	a	b	c
Chronic Obstructive Pulmonary Disease = a	0	3	0	3	0	0	3	0	0
Coronary Artery Disease = b	1	4	0	0	5	0	0	5	0
Pneumonia = c	0	0	8	0	0	8	0	0	8

Measurement - Logistic R (Cross-Val)	Prototype 1	Prototype 2	Prototype 3
Number of Attributes	21	24	146
Total Number of Instances	16	16	16
No: Correctly Classified Instances	7	8	8
No: Incorrectly Classified Instances	9	8	8
% Correctly Classified Instances	43.75%	50%	50%
% Incorrectly Classified Instances	56.25%	50%	50%
TP Rate Pneumonia	0.625	0.625	0.75
TP Rate Coronary Artery Disease	0.4	0.4	0.4
TP Rate Chronic Obstructive Pulmonary Disease	0	0.333	0
FP Rate Pneumonia	0.125	0.5	0.625
FP Rate Coronary Artery Disease	0.364	0.182	0.273
FP Rate Chronic Obstructive Pulmonary Disease	0.308	0.154	0
Precision Pneumonia	0.833	0.556	0.545
Precision Coronary Artery Disease	0.333	0.5	0.4
Precision Chronic Obstructive Pulmonary Disease	0	0.333	0
Recall Pneumonia	0.625	0.625	0.75
Recall Coronary Artery Disease	0.4	0.4	0.4
Recall Chronic Obstructive Pulmonary Disease	0	0.333	0

Confusion Matrix	Prototype 1			Prototype 2			Prototype 3		
Classified As	a	b	c	a	b	c	a	b	c
Chronic Obstructive Pulmonary Disease = a	0	2	1	1	1	1	0	1	2
Coronary Artery Disease = b	3	2	0	0	2	3	0	2	3
Pneumonia = c	1	2	5	2	1	5	0	2	6

Appendix T: Ghana Verbal Autopsy Sample results from WEKA

Story of Illness:

Ghana Verbal Autopsy (soi) training	OneR	ZeroR	J-Rip
Number of Attributes	51	51	51
Total Number of Instances	5	5	5
No: Correctly Classified Instances	2	2	2
No: Incorrectly Classified Instances	3	3	3
% Correctly Classified Instances	40%	40%	40%
% Incorrectly Classified Instances	60%	60%	60%
TP Rate Unexplained	0	0	0
TP Rate Severe Infection	1	1	1
TP Rate Congenital Abnormality	0	0	0
TP Rate Premature	0	0	0
FP Rate Unexplained	0	0	0
FP Rate Severe Infection	1	1	1
FP Rate Congenital Abnormality	0	0	0
FP Rate Premature	0	0	0
Precision Unexplained	0	0	0
Precision Severe Infection	0.4	0.4	0.4
Precision Congenital Abnormality	0	0	0
Precision Premature	0	0	0
Recall Unexplained	0	0	0
Recall Severe Infection	1	1	1
Recall Congenital Abnormality	0	0	0
Recall Premature	0	0	0
Confusion Matrix			
Classified As	a b c d	a b c d	a b c d
Unexplained = a	0 1 0 0	0 1 0 0	0 1 0 0
Severe Infection = b	0 2 0 0	0 2 0 0	0 2 0 0
Congenital Abnormality = c	0 1 0 0	0 1 0 0	0 1 0 0
Premature = d	0 1 0 0	0 1 0 0	0 1 0 0

Story of Illness continued.

Ghana Verbal Autopsy (soi) training	J48	NB	MLP	Adaboost	Log R
Number of Attributes	51	51	51	51	51
Total Number of Instances	5	5	5	5	5
No: Correctly Classified Instances	3	5	5	3	5
No: Incorrectly Classified Instances	2	0	0	2	0
% Correctly Classified Instances	60%	100%	100%	60%	100%
% Incorrectly Classified Instances	40%	0%	0%	40%	0%
TP Rate Unexplained	1	1	1	1	1
TP Rate Severe Infection	1	1	1	1	1
TP Rate Congenital Abnormality	0	1	1	0	1
TP Rate Premature	0	1	1	0	1
FP Rate Unexplained	0.25	0	0	0.25	0
FP Rate Severe Infection	0.333	0	0	0.333	0
FP Rate Congenital Abnormality	0	0	0	0	0
FP Rate Premature	0	0	0	0	0
Precision Unexplained	0.5	1	1	0.5	1
Precision Severe Infection	0.667	1	1	0.667	1
Precision Congenital Abnormality	0	1	1	0	1
Precision Premature	0	1	1	0	1
Recall Unexplained	1	1	1	1	1
Recall Severe Infection	1	1	1	1	1
Recall Congenital Abnormality	0	1	1	0	1
Recall Premature	0	1	1	0	1
Confusion Matrix					
Classified As	a b c d	a b c d	a b c d	a b c d	a b c d
Unexplained = a	1 0 0 0	1 0 0 0	1 0 0 0	0 1 0 0	1 0 0 0
Severe Infection = b	0 2 0 0	0 2 0 0	0 2 0 0	0 2 0 0	0 2 0 0
Congenital Abnormality = c	0 1 0 0	0 0 1 0	0 0 1 0	0 1 0 0	0 0 1 0
Premature = d	1 0 0 0	0 0 0 1	0 0 0 1	0 1 0 0	0 0 0 1

Csv File

Ghana Verbal Autopsy (csv) training	OneR	ZeroR	J-Rip
Number of Attributes	234	234	234
Total Number of Instances	2	5	5
No: Correctly Classified Instances	3	2	2
No: Incorrectly Classified Instances	2	3	3
% Correctly Classified Instances	40%	40%	40%
% Incorrectly Classified Instances	60%	60%	60%
TP Rate Unexplained	0	0	0
TP Rate Severe Infection	1	1	1
TP Rate Congenital Abnomality	0	0	0
TP Rate Premature	0	0	0
FP Rate Unexplained	0	0	0
FP Rate Severe Infection	1	1	1
FP Rate Congenital Abnomality	0	0	0
FP Rate Premature	0	0	0
Precision Unexplained	0	0	0
Precision Severe Infection	0.4	0.4	0.4
Precision Congenital Abnomality	0	0	0
Precision Premature	0	0	0
Recall Unexplained	0	0	0
Recall Severe Infection	1	1	1
Recall Congenital Abnomality	0	0	0
Recall Premature	0	0	0
Confusion Matrix			
Classified As	a b c d	a b c d	a b c d
Unexplained = a	0 1 0 0	0 1 0 0	0 1 0 0
Severe Infection = b	0 2 0 0	0 2 0 0	0 2 0 0
Congenital Abnomality = c	0 1 0 0	0 1 0 0	0 1 0 0
Premature = d	0 1 0 0	0 1 0 0	0 1 0 0

Ghana Csv continued.

Ghana Verbal Autopsy (csv) training	OneR	ZeroR	J-Rip
Number of Attributes	234	234	234
Total Number of Instances	2	5	5
No: Correctly Classified Instances	3	2	2
No: Incorrectly Classified Instances	2	3	3
% Correctly Classified Instances	40%	40%	40%
% Incorrectly Classified Instances	60%	60%	60%
TP Rate Unexplained	0	0	0
TP Rate Severe Infection	1	1	1
TP Rate Congenital Abnomality	0	0	0
TP Rate Premature	0	0	0
FP Rate Unexplained	0	0	0
FP Rate Severe Infection	1	1	1
FP Rate Congenital Abnomality	0	0	0
FP Rate Premature	0	0	0
Precision Unexplained	0	0	0
Precision Severe Infection	0.4	0.4	0.4
Precision Congenital Abnomality	0	0	0
Precision Premature	0	0	0
Recall Unexplained	0	0	0
Recall Severe Infection	1	1	1
Recall Congenital Abnomality	0	0	0
Recall Premature	0	0	0
Confusion Matrix			
Classified As	a b c d	a b c d	a b c d
Unexplained = a	0 1 0 0	0 1 0 0	0 1 0 0
Severe Infection = b	0 2 0 0	0 2 0 0	0 2 0 0
Congenital Abnomality = c	0 1 0 0	0 1 0 0	0 1 0 0
Premature = d	0 1 0 0	0 1 0 0	0 1 0 0

Appendix U: IHME Verbal Autopsy Sample results from WEKA

Measurement - Cross Validation	IHME ZeroR	IHME OneR	IHME JRip
Number of Attributes	132	132	132
Total Number of Instances	1592	1592	1592
No: Correctly Classified Instances	186	303	440
No: Incorrectly Classified Instances	1406	1289	1152
% Correctly Classified Instances	11.6834%	19.0327%	27.6382%
% Incorrectly Classified Instances	88.3166%	80.9673%	72.3618%
TP Rate Weighted	0.117	0.19	0.276
FP Rate Weighted	0.117	0.103	0.085
Precision Weighted	0.014	0.08	0.293
Recall Weighted	0.117	0.19	0.276

Measurement - Cross Validation	IHME J48	IHME Naïve Bayes	IHME MLP	IHME Adaboost
Number of Attributes	132	132	132	132
Total Number of Instances	1592	1592	1592	1592
No: Correctly Classified Instances	428	88	209	287
No: Incorrectly Classified Instances	1164	1504	1383	1305
% Correctly Classified Instances	26.8844%	5.5276%	13.1281%	18.0276%
% Incorrectly Classified Instances	73.1156%	94.4724%	86.8719%	81.9724%
TP Rate Weighted	0.269	0.055	0.31	0.18
FP Rate Weighted	0.05	0.021	0.49	0.106
Precision Weighted	0.244	0.132	0.127	0.057
Recall Weighted	0.269	0.06	0.131	0.18

Appendix V continued: IHME Verbal Autopsy Sample results from WEKA “x6” and “x16”

Measurement Cause of Death X6/X16	J48	Naïve Bayes	MultiLayer Perceptron	AdaboostM1	LogisticR
Number of Attributes	305	305	305	305	305
Total Number of Instances	143	143	143	143	143
No: Correctly Classified Instances	98.0328%	85.9016%	99.0164%	98.6885%	100%
No: Incorrectly Classified Instances	196.7200%	1409.8400%	98.3600%	1.3115%	0%
% Correctly Classified Instances	299	262	302	301	305
% Incorrectly Classified Instances	6	43	3	4	0
TP Rate x6	0.992	0.966	1	1	1
TP Rate x16	0.973	0.79	0.984	0.978	1
FP Rate x6	0.027	0.21	0.016	0.022	0
FP Rate x16	0.008	0.034	0	0	0
Precision x6	0.959	0.747	0.975	0.967	1
Precision x16	0.995	0.974	1	1	1
Recall x6	0.992	0.966	1	1	1
Recall x16	0.973	0.79	0.984	0.978	1
Confusion Matrix	**J48**	**Naïve Bayes**	**MultiLayer Perceptron**	**AdaboostM1**	**LogisticR**
Classified As	a b	a b	a b	a b	a b
x16 = a	181 5	147 39	183 3	182 4	186 0
x6 = b	1 118	4 115	0 119	0 119	0 119
Measurement Cause of Death X6/X16	**J48**	**Naïve Bayes**	**MultiLayer Perceptron**	**AdaboostM1**	**LogisticR**
Number of Attributes	305	305	305	305	305
Total Number of Instances	143	143	143	143	143
No: Correctly Classified Instances	281	250	282	294	248
No: Incorrectly Classified Instances	24	55	23	11	57
% Correctly Classified Instances	92.7311%	81.9672%	92.459%	96.3934%	91.3115%
% Incorrectly Classified Instances	7.8689%	18.0328%	7.541%	3.6066%	18.6885%
TP Rate x6	0.899	0.916	0.916	1	0.832
TP Rate x16	0.935	0.758	0.93	0.941	0.801
FP Rate x6	0.65	0.242	0.07	0.059	0.199
FP Rate x16	0.101	0.146	0.084	0	0.168
Precision x6	0.899	0.708	0.893	0.915	0.728
Precision x16	0.935	0.934	0.945	1	0.882
Recall x6	0.899	0.916	0.916	1	0.832
Recall x16	0.935	0.758	0.93	0.941	0.801
Confusion Matrix	**J48**	**Naïve Bayes**	**MultiLayer Perceptron**	**AdaboostM1**	**LogisticR**
Classified As	a b	a b	a b	a b	a b
x16 = a	181 5	141 45	173 13	175 11	149 37
x6 = b	1 118	10 109	10 109	0 119	20 99