Data Mining Occurrences of Infectious Diseases with SNOMED CT

By

Ewelina Ciolko

A Thesis Submitted in Partial Fulfilment of the

Requirements of the Degree

Master of Health Sciences

In

The Faculty of Health Sciences

Graduate Studies Program

University of Ontario Institute of Technology

May 2013

Abstract

Synonyms within SNOMED CT's structure give meaning to the clinical terminology. The hypothesis in this thesis is that the number of synonyms of a disease within SNOMED CT can be used to predict the number of occurrences of an infectious disease reported on by the World Health Organization (WHO). Using simple Classification and Regression (CART), Bayes theory, and Best Fit trees, prediction algorithms are created based on the number of synonyms in infectious disease terms of SNOMED CT, the number of those diseases world-wide, the region of occurrence of the disease, and the year of occurrence of the disease. The results of experiments predict the number of occurrences of a disease correctly 67% of the time by using Simple Cart method; Bayes and Best Fit Trees each produce the correct number of occurrences 61% of the time.

Keywords: SNOMED CT, data mining, World Health Organization, infectious diseases, simple CART theory, Naïve Bayes, Best Fit Trees, World Health Statistics

Acknowledgements

There are many people without whom this thesis would not be possible. Firstly I would like to acknowledge the Professors involved in this thesis: Dr. Fletcher Lu and Dr. Carolyn McGregor. They have helped to direct this thesis in focussed manner and have been supportive in its completion. I would also like to acknowledge Dr. Andrew James for providing a valuable perspective on SNOMED CT and how my research could relate to the medical world.

Secondly I would like to acknowledge my family for continuing to support me being a perpetual student. Thank you for your support mom and dad and for rooting for me and believing in me. Emily thank you for dealing with your “always a student” sister and for joining me for relaxing breaks during my studies. It is so important to have love in ones’ life, especially when faced with difficult tasks and I would like to acknowledge my family for providing that for me.

Last but not least, I would like to acknowledge Jamieson Churchill, my fiancé, for constantly wanting to stay in with me and do homework. It’s so much easier to do work when you have someone there to help motivate you and bounce your thoughts off of. Thanks so much for always being there with love and patience.

Table of Contents

List of Appendices

List of Figures

List of Tables

1 Introduction

An initial description of SNOMED CT will be provided in this section. In section 1.2 the objectives and methods of the thesis are described. The next sections of the introduction describe the research aims, and contributions. Section 1.5 provides an overview of the structure of the thesis document itself by listing the chapters and by providing a short description of each chapter.

1.1 Background/Rationale

This thesis uses data mining techniques to ascertain if the number of synonyms or nuances used to describe infectious diseases in SNOMED CT, is correlated with the number of occurrences of that infectious disease. I reason that since SNOMED CT is created based on a need and use basis, concepts in SNOMED CT should be better described if they are used more commonly. SNOMED CT can be considered a medical natural language lexicon which has been created based on need. This section introduces this concept by describing research previously completed with SNOMED CT and provides a background on the development of SNOMED CT.

Research which has used the term “ontology” to describe SNOMED CT will be included in this thesis. For this reason I will also sometimes refer to SNOMED CT as an ontology for consistency. SNOMED CT is sometimes referred to as a domain ontology made up of many domain ontologies. A domain ontology is one which specifies the concepts and relationships between concepts, in a particular area created by a team of experts (Boyce, Pahl, 2007).

SNOMED CT has been used to research medical burdens of diseases. Simpson et.al (2007) have used SNOMED CT in order to improve their understanding of the burden of allergies. They also made some recommendations to SNOMED CT about fixing codes located in the allergies

section of SNOMED CT. This research proves that SNOMED CT can be used for more than just retrieval of information.

Heja (2008) suggests that there is a lot of noise within SNOMED CT due to the method of its creation. This thesis suggests that the method of its development is the most valuable part of SNOMED CT. According to the International Health Terminology Standards Development Organization (IHTSDO) SNOMED CT is updated by members of IHTSDO who join working groups. In some research SNOMED CT has been referred to as a living breathing ontology which is constantly being updated based on its real world use (Heja, 2008). In this thesis I test this statement by testing if the quantity of descriptors of a disease is positively correlated to the number of occurrences of that disease. I would argue that most synonyms and nuances in SNOMED CT are reflective of real life as it is developed by clinicians who improve it as they use it.

There is research that shows that SNOMED CT is not useful for computations because it contains many complex relationships which have meaning and are more than just “is a” (Goldfain, Cowell, Smith, 2010). In this research I suggest that SNOMED CT can actually be used in computational research and that its value lies in the fact that it is created based on need. By focusing on non-determinant relationships within SNOMED CT that do not fall into the “is-a” category, I will test if the number of ways to describe a disease in SNOMED CT is related to the number of times that disease occurs.

1.2 Objectives/Methods

The objective of this thesis is to use data mining techniques to ascertain if the number of synonyms or nuances used to describe infectious diseases in SNOMED CT is correlated with the number of occurrences of that infectious disease. This will be met in this thesis by providing a background overview of the structure of SNOMED CT code and how it is used in health informatics today. Then three machine learning methods are selected based on the need to determine if the number of terms used to describe infectious diseases in SNOMED CT and the number of occurrences of that disease are positively correlated. After this the experiments are conducted and results are discussed. A contribution in the area of SNOMED CT development and machine learning in health care are made.

I use a constructive research approach to find aspects of SNOMED CT to use on data mining algorithms to predict real medical outcomes on an international level. “A constructive research method implies building of an artifact (practical, theoretical or both) that solves a domain specific problem in order to create knowledge about how the problem can be solved (or understood, explained or modeled) in principle” (Lukka, 2003). This thesis uses a constructive approach by first finding a domain specific problem: using SNOMED CT for more than data storage and a repository of medical terms. This problem is used to create a data mining algorithm, the practical artifact, to help find an approach to solving the problem..

1.3 Research Aim/Research Questions

The main research aim is to apply a machine learning algorithm to SNOMED CT in such a way that it will support the terminology’s usability and value in a novel way. In order to accurately assess this, the research uses the data on infectious disease occurrences from the World Health Statistics Reports published by the WHO.

The research questions, which will be answered in this thesis, are:

1. Can the number of synonyms and nuances in SNOMED CT's description of certain infectious diseases outlined by the WHO be used in order to develop an algorithm that helps predict the number of occurrences of that infectious disease?

2. Which data mining method commonly used in health informatics research can provide the best outcome when predicting the number of occurrences of an infectious disease?

1.4 Contributions

The main contributions of this thesis are:

- Providing evidence which suggests that SNOMED CT develops in correlation to how often its terms need to be used
- Demonstrating by exploration that parts of SNOMED CT's structure can be used, in part with other health data, for prediction purposes

1.5 Thesis Organization

This thesis consists of seven chapters. Chapter 1 provides an introduction to this work and the rest is as follows:

Chapter 2 introduces SNOMED CT and describes its structure and ontological aspects. This chapter also includes a literature search of how the ontological structure of SNOMED CT is used for prediction purposes in health care.

Chapter 3 presents the different data mining methods which were considered for this work. Each of their respective contributions to health care is introduced. This chapter also includes an assessment of which machine learning methods would be most suitable for health care data organized like SNOMED CT.

Chapter 4 introduces the World Health Statistics Reports which are used in this work. These statistical reports provide the numbers of infectious disease occurrences on an international scale. The SNOMED CT terms used are based on the infectious disease statistics available in these reports.

Chapter 5, the Methods chapter, describes the tools used in this research, as well as the different machine learning methods: Simple Cart, Bayesian, and Best Fit Trees. It also described the SNOMED browser, SNOB, which was used to gather the SNOMED CT codes. This chapter also describes all of the work that was done in the WEKA Explorer tool.

Chapter 6 provides the results for the initial correlation work done to show that SNOMED CT synonyms are positively correlated with the occurrences of infectious diseases reported by the WHO. It also provides all of the results for the data mining work done in WEKA. The results are then assessed by the researcher. The reliability and validity of the work are assessed in this chapter as well.

Chapter 7 presents the future recommendations for research in this field the outcomes of the research and a personal reflection on research in health sciences.

2 SNOMED CT

This chapter will describe SNOMED CT by first explaining how it is structured in Section 2.1. In Section 2.2 a literature search of the terms “ontology”, “data model representation”, “data mining” and “SNOMED CT” is performed. It is important to perform this literature search in order to understand the research which has been performed with SNOMED CT in the artificial intelligence (AI) field.

The review performed within this chapter enabled the quantification of the following research question within this thesis:

> Can the number of synonyms and nuances in SNOMED CT’s description of certain infectious diseases outlined by the WHO be used as part of an algorithm which helps predict the number of occurrences of that infectious disease?

2.1 Structure of SNOMED CT Code

This chapter will introduce SNOMED CT by describing how it is structured. This relates to the research question because in order to understand how to perform any data mining on SNOMED CT, it is important to know the potential aspects of it which could provide a real contribution.

SNOMED CT is a standardized clinical terminology system. It includes a “comprehensive coverage of diseases, clinical findings, therapies, procedures and outcomes. It provides core terminologies for an electronic health record” (Elevitch, 2005).

It contains over 366,000 concepts and over 993,000 English language descriptions of synonyms. This creates a reliable search engine for physicians to perform queries related to pathology of diseases (Elevitch, 2005).

Although SNOMED CT is comprehensive on its own it also maps to other medical classifications and terminologies which are widely used in the medical world (Elevitch, 2005). Among these is ICD-9. SNOMED CT ensures that the codes do not overlap and at the same time allow ICD-9 codes to represent the billing and administrative parts of the patients journey (Elevitch, 2005).

SNOMED CT has been proven to have a sensitivity of 92.3% and a positive predictive value of 99.8% (Elkin, et.al, 2006). Sensitivity in the context of that research is when SNOMED CT correctly assigns a code to a clinical term. Positive predictive value in the context of that research is when SNOMED CT ties certain clinical terms to other terms correctly.

SNOMED CT is owned and maintain by the International Health Terminology Standards Organization (IHTSDO). It is used by over 18 countries as of January 2012. All of the clinicians in these countries have the opportunity to submit corrections to SNOMED CT and do analysis on different parts of the clinical terminology. This provides SNOMED CT users with data that is meaningful because it is created in response to a real need to represent diseases properly in the field.

The second way to see that SNOMED CT is meaningful is to look into its structure. SNOMED CT represents medical data in a semantic way. The structure of SNOMED CT provides a look into the semantic nature of medical data as it happens in the real world. Many researchers found that SNOMED CT proved to have too difficult of a structure to be understood with automated methods.

SNOMED CT code identifies medical data as concepts. It distributes information into concept tables. There is also a description included which is a readable term. Each concept may

have more than one related description which covers all different ways of stating one concept. There are also synonyms included for most concepts to increase the usability of the terminology. Each description is represented by a code and is organized in a description table (Donnely, 2006).

Each relationship has an ID and is also distributed into a relationship table. A relationship identifies two related concepts and the type of relationship they have. Relationships also have a relationship group, characteristic type, and also devices which may have been used in the medical intervention (Donnely, 2006). The information from each table is organized into a hierarchy (Figure 1) which is later used when the data is called upon by a physician (Donnely, 2006).

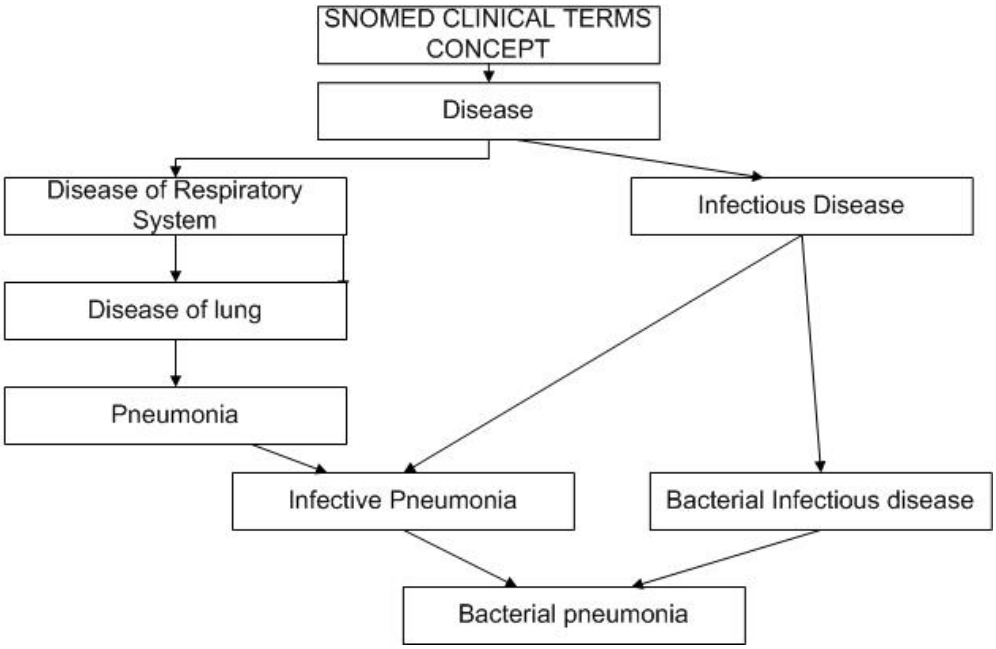

Figure 1 Graphical Representation of the "is a" relationship of an example concept (Donnely, 2006)

There are so many clinical terminologies existing in the medical world that even though SNOMED CT is highly

recommended, it is not always the choice selection of the medical facility and the databases that do exist are not always complete. This is unfortunate because it means that in the future there will be medical databases containing large amounts of data which would be used in medical research, but it may be in different terminologies, making it more difficult to use in data mining. If SNOMED CT could be shown to be of more value than just for storage, it is possible that the medical world would be more receptive to adapting to its use. Using SNOMED CT in a new way would increase its value and promote the research into new uses for clinical terminologies.

2.2 SNOMED CT Ontology-Literature Search

The IEEE XPlore database was used to search for literature for this review. The terms searched were as follows: “ontology”, “data model representation”, “data mining” “mapping”, “decision support” and “SNOMED CT” in a full text and metadata search. These terms were searched because they are the ones which are determined to be the most relevant based on the keyword search to being able to determine that SNOMED CT can be used for more than just data organization. The relevancy of these articles was determined by the number of keywords which were included. Ideally I aimed to find articles which included all the keywords. The search was limited to conference papers and journal papers dating from 1996-2010. This resulted in over 4000 articles being found. This large number of results displays the popularity of ontologies being used in databases. After further analysis of the results, only the first fifteen articles which were relevant to the topic of this literature search were chosen because they were the only ones to include two or more of the following topics at the same time: medical ontology, SNOMED CT, or prediction in health care.

Table 1 Comprehensive Literature Search

Terms Searched	Database,Search constraints and Keywords	Number of articles returned	Number of relevant articles	Titles of useful Articles	Theme and Relevance
Ontology in decision support	CINAHL via Ebsco host 2005-2010English languageJournal subset: computer/information scienceSpecial interest: evidence based practice	67	1	1. Using AI to bring evidence based medicine a step closer to making the individual difference	n/a
SNOMED ontology decision support	CINAHL via ebsco host Search within full text of articles	0	0		n/a
SNOMED ontology		0	0		n/a
SNOMED mapping decision support	Pubmed Search full text	3	1	2. Biomedical ontologies in action: Role in knowledge management, and decision support	n/a
SNOMED ontology	Pubmed Search full text	25	7	3. Formal representation SNOMED CT expressions 4. Ontological analysis of SNOMED CT 5. Auditing semantic completeness of SNOMED CT 6. Auditing relationships using a converse abstraction network 7. Comparison of ontology based semantic similarity measures 8. Practical impact of ontologies on biomedical informatics	n/a
PUBMED	References from other articles Keywords: terminology, quality control, information systems	1	1	9. Quality assurance of medical ontologies	Quality of medical ontologies..tools such as OntoClean:.barriers to assuring content quality of an ontology
Google scholar	Reference from earlier article	0	1	10. Title of book: Clinical decision support: the road ahead Chapter title: ontologies, vocabularies, and data models	Why patient data should be coded rather than natural language

2.2.1 Results

It can be seen in the results of Table 1 of this literature search that ontologies are widely used in the area of data mining, web mining and data preprocessing with the "Ontology in decision support" search term returning 67 articles. The literature presents many different uses for ontologies, but all with the same general purpose: to allow for quicker, easier, more effective data mining and workflow. The literature found also appears to separate the use of ontologies into categories which will be discussed in the next section of this literature review. It can be seen in this literature that ontologies have not been used directly as data for data mining, and neither has SNOMED CT. Although generally widely used, ontologies are only used in data mining for their structure, and not as a data source itself. This signifies that this thesis can add a new contribution in this area.

2.2.1.1 Web Mining

There are a number of articles which focus on the use of ontology to help with Web-Mining. The premise behind this section of research is that the World Wide Web has too much information for a user to be able to gather useful information (Tao, Li, Zhong, 2011). To deal with this problem many researchers have introduced the concept of personalized user ontologies which follow the needs of the user to help them find what they are looking for (Yuefeng, Zhong, 2006). This concept is sometimes known as Web Intelligence (Zhong, Liu, Yao, 2002). In this concept the user has certain behaviours and preferences which can be categorized into an ontology as they use the web for searches. For example a business owner may have different patterns and needs from online searches than a student, even though they search for the same terms (Tao, Li, Zhang, 2011).

According to the literature, ontologies are the best tool for developing user profiles which can be used for web mining. In these cases, web mining is seen as a typical web search prompted by a user inputting search terms, however with a more intuitive method for finding what the user may actually be searching for online. In order to create a basis for the ontology models, the researchers must first create a base of common terms. One way of doing this is by using "world knowledge". This is knowledge which is considered common sense to everyone and is "acquired through experience and education" (Zadeh, 2007). This study used a tool called Ontology Learning Environment which assisted users with extracting subjects of interest from the Web. It lets users signal which concepts are negative and which are positive. This project also tested their created ontology against other models and found that ontologies had the best recall rate with a proportional loss in precision (Tao, Li, Zhang, 2011).

Another way of creating a basis for knowledge is to check what the user's basic actions and choices are without paying attention to the implied knowledge. The way of doing this is to allow the users to search for web pages which they are interested in and then to document the "terms" or "primitive objects" which are found in this way. An algorithm is then created and testing is done to find patterns. Any patterns which are not correct are then marked as wrong and put back into the system until it can find only correct patterns. Although this seems like a tedious approach, it allows for an 81.48% recall rate (Yuefeng, Zhong, 2006).

Yet another use for web mining ontologies is for e-learning. Students are taking to online classes to do their work, however, not every student learns in the same fashion. In e-learning ontologies can help to manage the course material and to present it to students in a more personalized manner (Middleton, Shadbolt, DeRoure, 2004) (Alani, Dasmahapatra, O'Hara, Shadblot, 2003) (Keleberda, Lesna, Makovetskiy, Terziyan, 2004). In this study a new method of

creating ontologies is presented; researchers use a Bayesian approach for creating ontology to represent the student's preferences. A Bayesian model was created based on tests written by students and their results of these tests. Teachers were made to sketch out a general ontology based on how students learn and do on tests, and later a Bayesian model was mapped to this ontology to ensure its accuracy. This ontology was meant to help students in the learning process, and in fact it did. Students who used the tool did better on tests than students who did not (Colace, De Santo, 2010). This research presents a whole different aspect of ontology creation because it shows that it can be used to affect people's behaviours and help them succeed in daily tasks. It also shows us an ontology which can be verified using an artificial intelligence algorithm.

2.2.1.2 Ontology to Assist with Data Mining Decisions and Processes

The second category which came through in this literature review is that of using ontologies to help with the initiation of the data mining process itself. This means to help choose the data mining method as in Choinski & Chudziak (2009) and Bernstein, Provost & Hill (2005). These papers recognize the challenge faced by data mining experts in the ever growing field of data mining; data mining experts may have difficulty choosing the best possible data mining tool because it is difficult for them to know everything about the data which they are to mine (Ulrich, Eppinger, 2004). These papers suggest relatively the same solution to this problem: ontology based learning assistant or intelligent discovery assistant. In both cases, the assistants are meant to direct data mining experts to use the correct data mining process or method.

In Choinski & Chudziak (2009) the researchers aim to find a method of communicating between those individuals making business goals and those who are doing the technical work.

This research focuses on creating domain ontologies based on the CRISP-DM process, and then using these domains to help those working on knowledge data discovery projects. This idea is supposed to allow for all stakeholders in the data mining process to access and understand other areas of the process based on their specialties. A technology worker should be able to see and understand the business process part of the model based on her perspective and vice-versa. This is an ongoing project called Ontological Learning Assistant (OLA) which can be used in such cases as detecting when customers will leave cell phone or cable companies.

Bernstein, Provost, Hill (2005) help to understand the data mining process in mostly a technical way. It helps the data mining expert choose a data mining technique which will best suit the data in question. This research presents the idea of an Ontology-Based Intelligent Discovery System (IDA). This system would ask the user to "specify desired trade-offs between accuracy and speed of learning" (Bernstein, Provost, Hill, 2005). It would then suggest a data mining method to the user based on the user specifications. The ontology which was created in this research is based on three parts: pre-processing, induction algorithm, and post processing. In each case the ontology can be followed to determine the exact type of sampling method. The type is a strength and possibly a necessity to ensure that the resulting ontology is relevant to what the user needs.

Often times in the research, as in Colace & De Santo (2010) the process of making the ontology is tedious and should not take as long. In this study the teachers who were participants had to create ontology to represent the general process followed by a student who would do well in the course and students who would not. This proved to take up a lot of time and to be an extremely tedious task. For this reason the researchers decided to use a questionnaire filled out

by the students to automatically create ontology. This research later mapped a Bayesian network to this ontology to ensure that it was sufficient to use and to attach probabilities to the nodes..

2.2.2 Literature Discussion

The user profile ontologies were created in relatively the same way in these studies. The studies proved that ontologies can help in web mining by creating user profiles. They had relatively the same methods of creating and testing the ontologies. All of the methods included the actual user which is very different from other ontologies. Ontologies usually do not include probability features. This is possibly the direction in which the future of ontology creation is headed.

Using ontology to figure out which data mining method to use and how to use it, is a very important contribution to this area of research. This is a truly innovative use of ontology because up until now, ontology was meant to help in the data mining process, not in the selection of data mining tools. With the advent of terabytes of new data streaming in from all corners of the electronic world, it is important to have a standardized method of dealing with and learning from the data. The ontologies discussed in this literature review also both incorporate a business aspect to the data mining. This is especially important because data mining can be very costly and it is important for researchers on the technological side not to forget this. In order for the proposed data mining methods to be useful in real world situations, especially in small businesses or health care, they have to be cost effective. Using these ontologies which were created for choosing data mining processes and methods would be an intelligent thing to do rather than to not follow any standards, which could end up costing lots of wasted time and money.

2.2.3 Conclusions and Implications for Research

I propose that there are still other ways to use SNOMED CT in data mining which have not been used in the research literature just described. None of these articles describe research which uses parts of the ontology in prediction; the ontology is created as a prediction tool itself. This literature search determines that this thesis makes a contribution to this area because it shows that the research in this thesis has not been carried out before. This thesis contributes to research in SNOMED CT by showing that the data mining can be done in a different way than was previously done.

3 Machine Learning Techniques

In order to decide which machine learning technique would be most appropriate to use with SNOMED CT it is necessary to have a basic understanding of the different machine learning techniques.

- Decision trees (DT) Bayesian networks
- Support Vector Machines (SVMs)
- Gaussian processes (GPs)
- Simple Classification and Regression (simple CART)

This chapter provides the background to answer the research question: Which data mining method commonly used in health informatics research can provide the best outcome when predicting the number of occurrences of an infectious disease? It will do so by describing each data mining method and then describing various health informatics research applications which employed that data mining method. A comparison table is provided showing the most important points of the findings of this chapter.

3.1 Decision trees (DT) and Random forests

The algorithms used in DT learning search for the descriptive attribute that is chiefly related to a target variable. As the name states the output is a tree-shaped model that represents a small set of variables that together have a high predictive power for the target variable. If-then rules can also be used to describe the tree, as each rule can be identified by following the branches from the root node to a terminating node (Meyfroid, Guiza, Ramon, Bruynooghe, 2009).

Random forests (RF) are models that are a collection of decision trees (Figure 2). The result is a dataset with some duplicates and some examples left out. The RF model is a set of trees learned for each dataset and averaged over their predictions to get a final prediction. At times there are small deviations in a dataset that can have influence on a learned tree. The RF is a way of avoiding these unnecessary influences (Meyfroid, Guiza, Ramon, Bruynooghe, 2009).

3.1.1 DT and RF uses in healthcare

DT's have been found to be beneficial in Neonatal Intensive CareUnits as they are used to classify streams of received physiological signals and detect artifacts, thereby reducing the high number of false alarms. DT's are easy to understand and sharp in labeling errors and noise. This feature of DT's would be very useful in a Clinical Decision Support System because there are a number of possible errors that can be made in a patient's record.

It has been found that compared to DT, the RF has consistently been shown to perform better in the Intensive Care Units prediction tasks (Meyfroid, Guiza, Ramon, Bruynooghe, 2009).

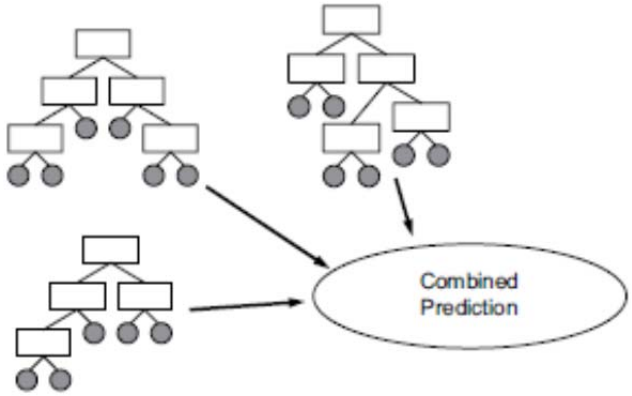

Figure 2 Random Forests (Meyfroid, Guiza, Ramon, Bruynooghe, 2009).

3.1.2 Bayesian Networks

A Bayesian network is a probabilistic graphical model that uses a set of random variables to specify a joint probability distribution. There are two fundamental components: a directed acyclic graph showing dependencies and independencies between variables, and a set of probability distribution tables. Within supervised learning there are two classes of Bayesian networks: Naïve Bayesian networks (NB) and Tree-Augmented Naïve Bayesian networks (TAN) (Figure 3). In NB the non-target variables are independent of other non-target variables; and there is a link from each non-target variable to the target variable. The TAN, in contrast, takes into account dependencies between non-target variables by having links between them (Meyfroid, Guiza, Ramon, Bruynooghe, 2009).

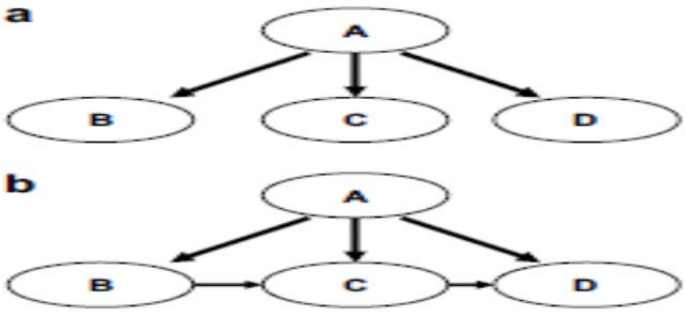

Figure 3 a) NB showing no dependencies between non-target variables B, C, D. b) TAN NB showing dependencies between the non-target variables B, C, D (Meyfroid, Guiza, Ramon, Bruynooghe, 2009).

Bayesian Networks in healthcare

Bayesian networks have been used in medicine since the 1970s; because they have been shown to successfully deal with uncertainties which are commonly present in clinical practice. Their robust capabilities allow the Bayesian models to be used in a wide range of complex domains such as medicine (Kabli, McCall, Herrmann, Ong, 2008).

3.1.3 Support Vector Machines (SVMs)

The origin of SVMs is from statistical learning theory and its essential component is the separating hyperplanes. A hyperplane is a geometrical division or separation between two outputs, such as predicting ICU mortality or survival (Meyfroid, Guiza, Ramon, Bruynooghe, 2009).

Support Vector Machines in healthcare

SVMs have been used for classification in medical domains. In one study a SVM was used to distinguish false signals from microcalcifications in digital mammograms. The SVM classifier performed slightly better than the Artificial Neural Network (ANN) counterpart, and was found to be easier to train (Meyfroid, Guiza, Ramon, Bruynooghe, 2009).

3.1.4 Gaussian Processes (GPs)

GPs give a prior probability to every possible function, with higher probabilities for the functions that are more likely. GPs allow for multi-dimensional inputs, have a small number of tunable parameters and result in full predictive distributions as opposed to the point predictions typical or other methods. GPs are found to consistently outperform more conventional methods such as ANNs in different regression tasks (Rasmussen, Williams, 2006).

Gaussian Processes in Healthcare

The use of GPs for regression has recently begun in the intensive care domain. GPs outperform other modeling methods for the analysis of electroencephalograph (EEG) signals to detect neonatal seizures. Another application of GPs has allowed the classification of patients according to the time frame in which they can be weaned from mechanical ventilation (Meyfroid, Guiza, Ramon, Bruynooghe, 2009).

3.1.5 Simple Classification and Regression Trees (simple CART)

Simple CART is a type of decision tree. It uses a pruning strategy that is different from other decision tree algorithms. The optimal tree is decided from all the possibilities available. There is no stopping in the pruning (Sathyadevi, 2011). CART moves down the tree and splits nodes that are not contributing to the optimal tree until every node contributes as much as possible. Then the optimal tree is selected.

Simple CART in Healthcare

There are a many areas in healthcare which are currently benefiting from Simple CART. In one case Simple CART was used to create a data mining approach for coronary artery disease screening (Pal, Chakraborty, Mandana, 2011). Another study uses Simple CART in a blood donors clinic. By using Simple CART they were able to identify a regular blood donor and in turn created a model which would help plan for more effective blood donation camps (Santhanam, Sundaram, 2010). Simple CART was also used in a study to help in the diagnosis of hepatitis. Previous patient diagnoses were entered into a Simple CART classification system and the outcome was found to have a good accuracy for predicting a diagnosis of hepatitis (Sundaram, 2010).

3.2 Comparison of Machine Learning Techniques

To decide which of these methods is best suited for our purpose we summarized previous research and previously made comparison of methods directly. This was achieved by assessing

the advantages and disadvantages in their use as shown by previous research in health care (Table 2).

Compared to other machine learning techniques, SVM's are known to be popular in predicting stock trends. Stock trends include large amounts of data coming in very often. This would suggest that SVM's are effective at data sets which have large amount of data coming in very often. They may not be the best fit for clinical practice which can have a smaller amount of data streaming in, but more for biological changes in the body (Chang, Fan, Dzan, 2009). This is possible because SVM's could be made to work best on biomedical data, rather than small sets of data with few data points spreading over large periods of time which can be true for clinical data.

In a recent study completed on the prediction of gastric cancer, a comparison was made between neural networks and decision tree models. Neural networks were shown to be better predictors of gastric cancer than the decision tree model (Wang, Li, Hu, Zhu, 2009).

A comparison study was performed on three different machine learning techniques:

Naïve Bayes, neural network, and decision trees. They were used to predict the likelihood of a patient developing heart disease based on medical profiles such as age, sex, blood pressure and blood sugar. The study showed high performance by all three models, but the Naïve Bayes model appeared to outperform the other two as it gave the highest number of correct predictions (Palaniappan, Awang, 2008).

Table 2 Comparison of Machine Learning Techniques

Type	Advantages	Disadvantages	Uses in Healthcare	SNOMED CT
Decision Tree (DT)	Easy to understand Robust in labeling errors & noise	Popular in labeling rather than producing new knowledge	Detects artifacts in streams of physiological signals	●
Random Forests (RF)	Doesn't get influenced by slight variations in data		Used for ICU data & prediction tasks	
Artificial Neural Networks (ANNs)	detects errors & learns from noisy examples Used in data analysis, pattern recognition, & prediction	Long training Time	Used in predicting gastric cancer	
Bayesian Network	Used in complex domains such as medicine to derive relationship probabilities and pathways		Used in medicine since 1970s Deals well with uncertainties commonly present in clinical practice	●
Support Vector Mechanism (SVMs)	Useful in logistic regression Used as a classifier, found to be better than ANN Easier to train	More popular in stock forecasting not medicine & cannot identify three-label data efficiently	Used for classification in medical domains	
Gaussian Processes (GPs)	Allows for multi-dimensional inputs Outperforms ANN in different regression tasks	Used more commonly for mapping physiological data streams	Recently used in intensive care domain Proficient in reading EEGs	
Simple CART	Form of regression Finds optimal decision tree Uses own pruning method leading to the most optimal possible solution Represented in easy to interpret visual form		Used to predict diagnosis and need for screening for Coronary Artery Disease Used to classify blood donors into types allowing for better blood donation camp scheduling	●

The purpose of this chapter was to show the different types of machine learning techniques that have been used in health care research before. By doing an analysis on these techniques a hypothesis can be made that Bayesian networks and Simple CART would be most suitable to assist in predicting the number of infectious diseases while also using aspects of SNOMED CT. This is a supposition made only from examining the literature and still requires experiments to verify. These experiments are described in Chapter 5.

4 The World Health Organization: World Health Statistics reports

This chapter provides the justification for using the data on infectious diseases provided by The World Health Statistics (WHO) and demonstrates how this is related to SNOMED CT. This is important to the thesis because it will introduce an idea of how SNOMED CT can be used for more than just storing data in a certain format. This chapter also describes where the data for numbers of infectious disease occurrences is taken from for this thesis. This is relevant to the thesis because it describes how the numbers are obtained by The World Health Organization and how accurate they are. This can affect the outcomes of the experiments by being a compounding factor so it is important to understand. If the number of infectious diseases is not accurate enough, then the question of can SNOMED CT be used to predict the number of infectious diseases cannot be properly assessed.

4.1 Why study data from The WHO?

The WHO releases a new report which includes statistics on infectious disease occurrences on a yearly basis (WHO, 2008). In Chapter 2 of this thesis it was also explained that SNOMED CT is released twice a year and is developed by members of the medical community who use it to represent diseases. An increase in disease occurrence can reasonably relate to an increase in SNOMED CT coverage of these diseases. The hypothesis being tested with the WHO statistics is that as the medical profession discovers more about a disease there should be more ways to describe it in a language that is meant to represent medical data on a world scale. This is an area of SNOMED CT which suggests that there is potential for data mining with SNOMED CT. If the

hypothesis that SNOMED CT coverage of terms is increased or improved as the disease represented by the term occurs more frequently in the real world then these two factors should have a correlation. In Chapter 5 Methods of this thesis's experiments to assess the correlation between these two factors are reported. This supports the overall research purpose of this thesis which is assess whether SNOMED CT can be used for purposes other than for organizing health data.

4.2 About the World Health Statistics reports

The data on the occurrences of the diseases and the regions they occur in comes from reports produced by the World Health Organization (WHO) called "World Health Statistics". These reports are a reliable source of information on the occurrences of infectious diseases around the world along with other global health indicators. The data found in these reports is used to make the WHO's yearly health-related Millennium Development Goals. This report has been made yearly since 2005.

The World Health Statistics report is "compiled using publications and databases produced and maintained by the technical programmes and regional offices of the WHO" (World Health Organization, 2008). The reports are also provided to the WHO using the International Classification of Diseases Versions 9 and 10 (WHO, 2008). The global health indicators mentioned in these annual reports provide a comprehensive summary of the current status of national health and health systems in nine areas. They are:

1. Life expectancy and mortality
2. Cause-specific mortality and morbidity
3. Selected infectious diseases

4. Health service coverage
5. Risk factors
6. Health workforce, infrastructure and essential medicines
7. Health expenditure
8. Health inequities
9. Demographic and socioeconomic statistics.

When interpreting the data, it is important to consider that certain countries have immunizations for certain infectious diseases while other countries do not. This influences the number of times the disease is reported in that country.

This data is not certified by the countries included in the reports creating some uncertainty with the data (WHO, 2008). Estimations in the data are made, and they depend on previous year's statistics and numerous mathematical methods which are not explained in detail by the WHO.

In conclusion it can be said that based on the above information, though the data provided by the World Health Organization may have some inaccuracies due to such factors as lack of data certification it is assumed in this thesis that it is sufficiently accurate for used in this thesis. The areas of the report which are not completely accurate may be compensated for by the data mining techniques described in Chapter 3. These techniques can account for noise and many other flaws which the data may contain. In the following Chapter 5 Methods, the data from the WHO statistical reports is used in order to test the hypothesis that SNOMED CT can be used for purposes other than organization of medical data.

5 Methods

This chapter provides an explanation of all the steps within the experimental design performed in this thesis. Section 5.1 describes the tools used in the experiments. It introduces the SNOMED browser used to find SNOMED CT synonyms and nuances and the data mining software, WEKA, used to perform the data mining portions of the experiments. Section 5.2 describes the data exploration completed in order to determine if there is a correlation between the number of synonyms in SNOMED CT and the number of occurrences of the infectious diseases based on the WHO statistics. Section5.3 shows all the work which was executed in WEKA explorer as well as how it was executed.

This chapter describes the experiments completed in order to answer the two following thesis questions:

1. Can the number of synonyms and nuances in SNOMED CT's description of certain infectious diseases outlined by the WHO be used in order to develop an algorithm that helps predict the number of occurrences of that infectious disease?
2. Which data mining method commonly used in health informatics research can provide the best outcome when predicting the number of occurrences of an infectious disease?

5.1 Tools

Two applications were used in this project. They are SNOMED Browser and WEKA explorer. These will be described in the next two sections.

5.1.1 SNOMED Browser (SNOB)

SNOB is a SNOMED CT browser available for free online. It is a tool that allows for one to see multiple editions of SNOMED CT in extensive detail. It is provided online by Eggbird. Eggbird is a Dutch company developing software in the area of medical terminology (http://www.eggbird.eu/).

There are other SNOMED CT browsers available online such as SNOMED CT Core Browser and CliniClue. SNOMED CT Core Browser did not include the opportunity to upload the browser onto my own personal computer (http://terminology.vetmed.vt.edu). Cliniclue provided this option, but did not provide the opportunity to see multiple versions of SNOMED CT at once. This was thought to be important for this thesis because if the synonym numbers were different between years it could easily be seen and later included in the analysis. Although this feature was not used in the thesis this was not known at the time of choosing the browser that it would not be used. Minnow is another SNOMED CT browser which was considered, however it did not include the ability for users to edit descriptions and registration was required. To choose a browser the U.S National Library of Medicine list of SNOMED CT browsers (http://www.nlm.nih.gov) was used.

5.1.2 WEKA Explorer

WEKA Explorer is a tool available to download online for free. It is software which contains many different machine learning algorithms. Users can download the software, then upload their data to it and execute an algorithm. WEKA provides the ability to process the data before applying an algorithm to it. There is nothing like WEKA freely available online anywhere else (Hall et.al, 2009).

WEKA is a project of the Machine Learning Group at the University of Waikato. Their team Vision is to "build state-of-the-art software for developing machine learning techniques and to apply them to real-world data mining problems (Hall et.al, 2009).

5.2 Data exploration

This chapter describes the steps done to complete the experiments conducted in this thesis. Figure 3 shows a flow diagram describing all the steps done.

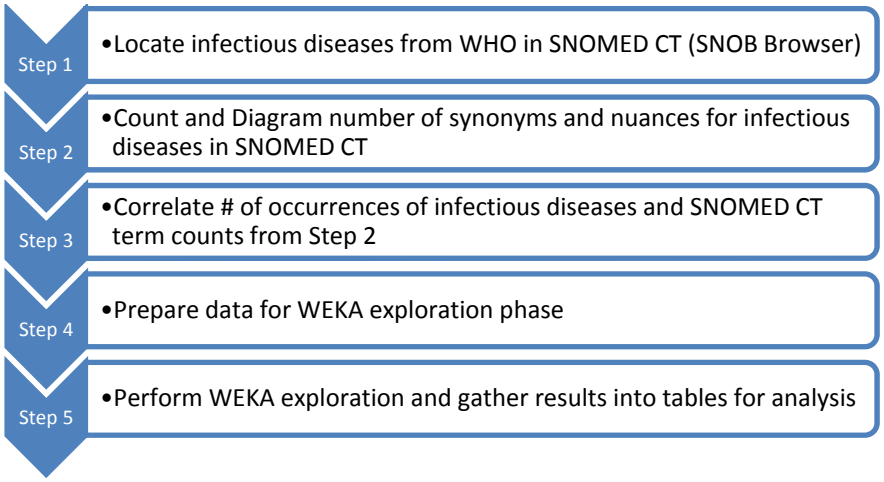

Figure 4 Steps in data exploration phase.

Step 1 was to locate the infectious diseases described in the WHO World Health Statistics reports in the SNOMED CT browser SNOB. These were located by executing a search of the database of terms in SNOB. In this part of the analysis a ledger of the SNOMED CT terms describing the infectious diseases as well as the number of ways to express that disease with the terminology was made (Appendix A). This count included synonyms for each disease as well as different categories of the disease. For example, there are 24 ways of expressing Rubella in SNOMED CT. This number includes synonyms and nuances all considered representative of rubella in its many forms. Table 3 shows all the terms included for Rubella.

Table 3 Rubella terms and Number Synonyms

Main Term	Synonyms	Variations
Gestational rubella syndrome	3	Congenital rubella, Congenital rubella syndrome, Gregg's syndrome
Congenital rubella pneumonitis	0	
Expanded rubella syndrome	2	Extended congenital rubella syndrome, ext. rubella syndrome
Progressive congenital rubella encephalomyelitis	0	
Rubella encephalomyelitis	0	
Rubella cataract	0	
Rubella	1	German measles
Endocochlear rubella	0	
Hemorrhagic rubella	0	
Rubella arthritis	1	Arthritis due to rubella
Rubella deafness	0	
Rubella in mother complicating pregnancy, childbirth AND/OR puerperium	0	
Rubella in pregnancy	0	
Maternal rubella in pregnancy, childbirth and the puerperium	0	
Maternal rubella duringpregnancy - baby delivered	0	
Maternal rubella during pregnancy –	0	
Maternal rubella in	0	
Maternal rubella in the puerperium - baby delivereddduring previous episode of care	0	
Rubella damage in pregnancy	0	
Rubella infection of central nervous system	0	
Progressive congenital rubella encephalomyelitis	0	
Rubella encephalomyelitis	0	
Progressive rubella panencephalitis	0	
Rubella meningoencephalitis	0	
Rubella with neurological complication	0	
Rubella myocarditis	0	
Rubella with complication	0	
Rubella without complication	0	

The number of synonyms and nuances for the infectious diseases were all calculated this way (Appendix A). Once this was complete, a comparison of these synonym counts to the number of occurrences of the infectious disease for each region described in the World Health Statistics report from the year 2010 was performed (Appendix B). Only one of the reports was included at this point to assess for possible data mining opportunities.

The results of this phase showed that there was no significant correlation between the number of synonyms and occurrences for each region. At this point it was necessary to analyze the data further in order to see if there was a cause for the lack of a correlation. It was noticed that the occurrences which had two synonyms describing the disease in SNOMED CT were outliers in every case. This prompted more exploration into why this could possibly be an outlier.

After viewing the infectious diseases which had two synonyms, it was found that all terms involved represented medical synonyms for diseases (Table 4). This is in contrast to the diseases which had more than two synonyms. The diseases represented by more than two synonyms were usually diseases which were broken down into more specific categories (Table 3) and diseases which were described with natural language nuances such as acronyms, as well as a synonym (Table 5). The conclusion drawn from this further exploration into two synonym terms is that the theory presented did not apply to terms with two synonyms because they were not related to how often the term is used, but rather to medical names. The terms with two synonyms were not dependant on natural language representation but rather on having different names describing the same thing and no input from the community clinicians. This means that the terms with two synonyms cannot be said to be dependent on how they are used in the community, this would exclude them from the theory which is being tested in this thesis. For this reason, the outlier is removed from the experiments.

Table 4 Diseases with two Synonyms

Disease	Number of synonyms	Terms
Malaria	2	Paludism, Plasmodiosis
Measles	2	Morbilli, Rubeola
Jungle Yellow Fever	2	Sylvatic Yellow Fever, Sylvan Yellow Fever

Table 5 Diseases with more than two Synonyms

Disease	Number of synonyms	terms
japanese encephalitis	4	JBE - Japanese B encephalitis JE - Japanese encephalitis Japanese B encephalitis Japanese encephalitis
Pertussis	3	Infection due to Bordetella pertussis WC - Whooping cough Whooping cough

The outlier was removed and again assessed for a correlation. This time there was a correlation found in each case (Table 4). Although there appears to be another outlier in the

graphs it was decided that the correlation values were high enough to leave the outlier in. Because this is a data mining algorithms experiment, it is acceptable to leave in such an outlier. This is because the algorithms have built in properties which deal with the outliers while in the learning phase. The data should be kept as real as possible before putting it in the WEKA Explorer so as not to skew the results away from a real word value. This is why only one outlier is eliminated at a time until a result with a valid correlation appearing to have promise for a real data mining result is reached.

The values calculated for each region from the 2010 WHO Report without the outlier of occurrences with two synonyms show correlation value that suggests there is a correlation (Table 6). This indicated the possibility that predictions could be made with this data, warranting further data mining.

Table 6 Correlation of Occurrences and SNOMED CT synonym numbers

Region	R^2 value
African Region	.6315
Region of Americas	.8265
South East Asia region	.7758
European Region	.9022
Eastern Mediterranean Region	.9269
Western Pacific	.9507

5.3 Work in WEKA Explorer

The next step was to use the data mining tool WEKA Explorer. With this tool data analysis which includes more dimensions than just the two already explored while finding a correlation can be done. For this the data has to be prepared to WEKA specifications which will be described in this chapter.

There are a total of four dimensions in the data. They are:

1. the number of SNOMED CT synonyms
2. number of occurrences from the WHO World Health Statistics reports
3. Region
4. Year

There were four different WHO reports included with data from the year 2007-2010 inclusive. In total there are 161 instances of data with four dimensions each. Four of the data points were excluded due to being outliers; this means there were 640 data instances entered into WEKA Explorer. One instance of data is made up of one number of occurrences, one date, one number of synonyms, and a region.

Before putting the data into WEKA, one file containing all of the data was created (Appendix C). The file type which is accepted by WEKA Explorer for classification purposes is the .arff format. This format separates the file into two sections. The first section lists and describes the attributes. In the first line of the file is the relation declaration. This line describes the title of the relation. In this study the title was @relation ‘diseases’. The next lines describe the attributes to be included in the data (Hall, et.al., 2009). The attributes must be listed in a specific order, as their order represents which column contains the data for that attribute.

In the data there are two types of attributes included: nominal and real attributes. The nominal attributes are ones which must be a certain value (Hall, et.al., 2009). All of the possible values are listed. For attribute “date” the possible values are listed as 2007, 2008, 2009 and 2010. The attribute “regions” is also categorized into nominal values. The values for regions had to be changed from words to numbers in order to comply with WEKA requirements (Appendix D). The other attributes, date and region are put in as real attributes.

The second section consists of the data itself. It is separated from the other section of the file with a row containing this header: “@data”. In order to be able to interpret the results of the software easily I have separated the numbers of occurrences categorized into ranges and represented by much smaller numbers (Appendix E).

After formatting the file I ran it through the classifier tool in WEKA Explorer. I chose three different algorithms: the Simple Cart Method, Best Fit Trees (a form of decision tree), and Naïve Bayes, based on the literature review of which algorithms would work best. Best Fit Trees were chosen based on the availability of decision tree algorithms in WEKA.

5.3.1 Test Parameters

For each test there were a number of parameters which could be selected. One of the configurable parameters was the amount of data to be used in the training phase for the algorithm and the amount of data to be used in the testing phase. We found that with different percentages of data used in the testing phase, the results changed. The best percentages are listed in Table 7.

Table 7 Percent of Data in Learning Phase

Algorithm	% of data used for learning phase
Naïve Bayes	66
Simple Cart Method	45
Best Fit Trees	45

5.4 Conclusion of methods

This section provided step by step instructions on how the experiments involved in this research process were conducted. All of the data preparation was explained. Based on the results of this work the following questions of this thesis can be answered.

1. Can the number of synonyms and nuances in SNOMED CT's description of certain infectious diseases outlined by the WHO be used in order to develop an algorithm that helps predict the number of occurrences of that infectious disease?

2. Which data mining method commonly used in health informatics research can provide the best outcome when predicting the number of occurrences of an infectious disease?

6 Results Findings/Discussion

This chapter presents the results of the experiments outlined in the previous chapter. Together these results present an additional usage for SNOMED CT within health care other than for storing health data in a meaningful way. The results of the correlation analysis between the rate of occurrences of a disease and the number of synonyms/nuances in SNOMED CT used to represent those diseases, are shown and explained first. After it is found that there is a positive correlation between these two variables and why this may be, the next step is to describe the results of the data mining experiments. The results of the data mining experiments show that it is possible to predict the number of diseases between 61%-68% of the time.

As a result this chapter provides the evidence to satisfy the following research questions as introduced in chapter one:

1. Can the number of synonyms and nuances in SNOMED CT's description of certain infectious diseases outlined by the WHO be used in order to develop an algorithm that helps predict the number of occurrences of that infectious disease?
2. Which data mining method commonly used in health informatics research can provide the best outcome when predicting the number of occurrences of an infectious disease?

6.1 Results of Correlation Analysis

The first tests in the experiments as detailed in Chapter 5 were to find if there was a correlation between the number of occurrences and the number of synonyms for the disease type

based on the World Health Statistics reports. This would justify the progression to the next stage of experimentation which was to perform data mining experiments on the data.

Table 8 Results of First Correlation

Region	Correlation value (r^2)
African	.0483
Region of the Americas	.0034
South East Asian Region	.0096
European Region	.0096
Eastern Mediterranean Region	.0454
Western Pacific Region	.2353

Table 8 illustrates the correlation value between SNOMED CT coverage of infectious disease terms and the occurrences in the population of certain regions. The first column describes the region tested. The second column displays the correlation value found when comparing the number of synonyms and nuances to the number of occurrences. This table includes an outlier associated with two SNOMED CT synonyms.

Table 9 Results of Second Correlation Analysis

Region	Correlation value (r^2)
African	.6315
Region of the Americas	.8265
South East Asian	.7758
European Region	.9022
Eastern Mediterranean Region	.9269
Western Pacific Region	.9507

Table 9 illustrates the correlation value of SNOMED CT synonym numbers and occurrences in the population of certain regions. The first column in this table describes the region of

occurrence of the disease as stated by the WHO report. The second column describes the correlation value found when comparing the number of synonyms to the number of occurrences in that region. This table takes out the outlier mentioned above. The outlier is an occurrence of disease in a population where the synonym or nuance number of terms in SNOMED CT for that disease is equal to two.

It can be seen that Table 9 presents good correlation values. These results show that there is potential for gaining some new knowledge by focussing on experimenting with these variables. The results for these experiments are shown in section 6.2.

6.2 Results of WEKA classification

Table 10 Results of WEKA Classification

	Simple cart method Actual/Percentage		Best Fit trees Actual/Percentage		Naïve Bayes Actual/Percentage	
Data Instances	89		55		55	
Correctly classified instances	60	67.4%	34	61.8%	34	61.8%
Kappa statistic	0.5815		0.5087		0.5135	
Mean absolute error	0.926		0.1005		0.1465	
Relative absolute error	45.5%		49.3%		71.9%	

Table 10 contains all of the results for the WEKA classification tests. The first row in the table describes the number of instances which the classifier was able to predict. Each of the results is displayed according to the classifier used and also shown in an actual value and a percentage. It can be seen that Simple CART method had the best recall rate. This is interesting because the same percentage of data (45%) to train the Simple CART method and the Best Fit Trees method was used. From these results it appears that Simple CART is the best algorithm to predict an outcome. This is also reflected in the values for all of the error calculations included in this chart; the relative absolute error at 45.5% are the lowest for Simple CART. The Kappa statistic is also highest for Simple CART algorithm. This does not necessarily mean that Simple CART is better at the prediction though, As Kappa values can be characterized as such: values < 0 as indicating no agreement and 0–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1 as almost perfect agreement (Landis, Koch, 1977). The Kappa statistic for all the methods show that the results are more than just chance outcomes, they have a moderate agreement.

The other two classifiers remain fairly similar to one another in most of the measures. One main difference was in the relative absolute error, where Naïve Bayes has a value of 82.0% BF trees have a value of 49.3%. The reason for this large difference in numbers is unknown; we can only assume that it is because the BF trees classifier is not as suitable for our data. The root relative squared error is higher for Naïve Bayes, but only by 7 points. This is not a large difference and should not be considered indicative of Naïve Bayes being a more suitable algorithm.

6.3 Reliability and Validity

A major issue with concerning the validity and reliability of the results stems from the WHO reports on World Health Statistics. As mentioned in Chapter 4 of this thesis, the data provided in the reports is not collected in a standardized way. Between countries there are differing methods of collection, different methods of reporting incidences, and numbers of occurrence of the disease can be influenced by factors such as national vaccination programs (World Health Organization, 2008). It is impossible to know to what extent the data is affected by these things.

This thesis is limited to using the synonym and nuance counts for infectious diseases which were reported on by the WHO. There are hundreds more diseases described by the infectious disease section of SNOMED CT. Infectious diseases are also just one medical subspecialty of SNOMED CT. For this reason it cannot be concluded that these results apply to all the infectious diseases in SNOMED CT. There are however certain expectations of generalizability for the outcomes of this thesis. The results of this work may very well when applied to other medical specialities such as internal medicine, surgery, obstetrics and gynaecology, paediatrics and psychiatry. Since all of these areas of SNOMED CT are created based on a needs basis it is possible that the correlation between synonyms and number of occurrences is positive. These experiments successfully use SNOMED CT in a novel way. The experiments show that SNOMED CT does not have to be limited to being a meaningful way to store data.

6.4 Implications for the domains of Computer Sciences, Health Informatics and Medicine

In this section the results of this thesis will be discussed from the perspective of the three domains present in this thesis: computer sciences, health informatics, and medicine.

The main implication of the results of this thesis to health informatics is to show that health

information is already stored and represented in a way that lends itself to data mining. It is important to discover new ways to use health information resources which already exist to help enrich knowledge about health. These results also prove that the when doing health informatics research one had to be aware of the nature of the data. There are many limitations to medical data. The reasons for limitations to the data can range from patient privacy regulations to the method in which the data is collected. It is helpful that terms are standardized across countries to allow for any comparison of the data at all. The importance of standardized methods of data collection and interpretation are highlighted by the fact that the number of correctly categorized occurrences of disease was not even at 70%. This is related to the implications of the results to the medical domain.

Clinicians are responsible for participating in the creation of SNOMED CT. The results of this thesis can highlight to medical professionals the importance of actively improving ways to represent medical data, especially clinical terminologies such as SNOMED CT. This thesis also highlights the importance of collection of medical data numbers from around the world. Standardization of reporting diseases is essential to make the data usable for research. The better the data and the more participation from physicians, the better the results of data mining will be. It is possible to create data mining tools which can help predict how many vaccines may be required in the next year, how many new cases of a disease there might be. This can only be possible if cooperation from the medical community is widespread.

From a computer science perspective, it is possible that there are data mining methods which can produce results which have a higher number of correctly classified instances. At 67% and 61.8% the number of correctly classified instances can be better. More exploratory research should be done in order to test algorithms which were not included due to not being used in similar

research before. The small number of data instances may be more suited to an algorithm which works better on small amounts of data as well.

7 Conclusions, Recommendations, and Personal Reflections

This Chapter will describe the final conclusions and outcomes of this thesis. It is concluded that the following two main research questions of this thesis have been fulfilled:

1. Can the number of synonyms and nuances in SNOMED CT's description of certain infectious diseases outlined by the WHO be used in order to develop an algorithm that helps predict the number of occurrences of that infectious disease?

2. Which data mining method commonly used in health informatics research can provide the best outcome when predicting the number of occurrences of an infectious disease?

Other than describing how these two questions were answered, this chapter will also describe limitations of this thesis, future recommendations will be made as well as a reflection on performing health informatics research.

7.1 Outcomes

This section describes the outcomes of the thesis experiments. The two thesis questions are answered in this section as well.

The results show that SNOMED CT can be used as more than for a standard data storage, expanding SNOMED CT's utility beyond a simple clinical terminology.

7.1.1 Summary

This thesis shows that with three different data mining methods it is possible to create desired outcomes while using an aspect of SNOMED CT's structure, the synonyms and nuances. Simple Cart, Naïve Bayes and Best Fit Trees were all found to have the ability to predict the number of occurrences of an infectious disease and answer the thesis questions. These data mining methods were chosen to be used in this thesis after a literature review was performed to discover how the methods were used in health care research in the past. Simple Cart was able to predict the number of occurrences correctly 67.4% of the time, Best Fit Trees and Naïve Bayes both predicted the number of occurrences 61.8% of the time.

The answers to the thesis questions are as follows:

1. Can the number of synonyms and nuances in SNOMED CT's description of certain infectious diseases outlined by the WHO be used in order to develop an algorithm that helps predict the number of occurrences of that infectious disease?

Yes the number of synonyms and nuances in SNOMED CT's description of certain infectious diseases outlined by the WHO can be used to develop an algorithm helping to predict the number of occurrences of the corresponding disease.

2. Which data mining method commonly used in health informatics research can provide the best outcome when predicting the number of occurrences of an infectious disease?

Simple CART, Naïve Bayes and Best Fit Trees were found to provide the best outcome when predicting the number of occurrences of an infectious disease.

7.2 Limitations and Complications

The main limitation and complication in this research was finding medical data to use. This was difficult to do because real medical data is protected by custodians who adhere to the Personal Health Information Protection Act . It is rare to find medical data which represents real data from real patients. Because the WHO had readily available information regarding occurrences of infectious diseases , this was the data used. This helped to shape the research question because the area of SNOMED CT which was going to be examined was decided by the type of medical data available.

7.3 Future Research Recommendations

As there have been improvements made to WEKA Explorer and other WEKA tools since these experiments were done, it is possible that there are more classifiers available to try on this data. There is the possibility that one of these classifiers may present better results than our Simple CART method and this could be worth exploring.

Since this research was performed there was also one more report completed and published on World Health Statistics by the WHO, adding these statistics should be a priority in future research as well. Since this research was finalized there have also been updates to various parts of SNOMED CT. It is possible that the synonyms for some infectious diseases have been added or taken away. Exploring these would be worthwhile as well to improve the system.

The next step for SNOMED CT research in this area would be to apply the same tests, using the synonym/nuance count, but from another category in SNOMED CT, such as Chronic Diseases, to see if the same rules apply. It would be beneficial to know if these results are

representative of what all of the synonyms/nuances in SNOMED CT can do, or only the synonyms/nuances for terms which describe infectious diseases.

8 References

Alani, H., Dasmahapatra , S., O'Hara, K. & Shadblot, N. (2003). ONTOCOPI-Using ontology based network analysis to identify network of practice, *IEEE Intelligent Systems*, 18 (2), 18-25.

Bernstein, A. &Provost, F.; Hill, S. (2005) Toward intelligent assistance for a data mining process: an ontology-based approach for cost-sensitive classification. *Knowledge and Data Engineering, IEEE Transactions*, 17 (4), 503-518.

Boyce, S., & Pahl, C. (2007). Developing Domain Ontologies for Course Content. *Educational Technology & Society*, 10 (3*)*, 275-288.

Chang, P., Fan, C. & Dzan, W. (2009). A CBR-based fuzzy decision tree approach for database classification. *Expert Systems with Applications*. 37, 214-225.

Choinski, M. & Chudziak, J.A. (2009) "Ontological Learning Assistant for Knowledge Discovery and Data Mining. *Computer Science and Information Technology*. 12 (14), 147-155.

Colace, F. & De Santo, M. (2010). Ontology for E-Learning: A Bayesian Approach. IEEE Trans. Education . 53(2), 223-233.

Donnelly, K. (2006). SNOMED CT: The advanced terminology and coding system for eHealth. *Studies in Health Technology and Informatics*. 121, 279-290.

Elevitch, F. (2005). SNOMED CT: electronic health record enhances anesthesia patient safety. *American Association of Nurse Anesthetists*. 73, 361- 366.

Elkin, P., Brown, S., Husser, C., Baver, B., Wahner-Roedler,D.,Rosenbloom,T. & Speroff,T.(2006). Evaluation of the content of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. *Mayo Clinic Proceedings*. 81, 741-748.

Goldfain, A.G., Smith, B. & Cowell, L.G. (2010). Dispositions and the Infectious Disease Ontology. in Antony Galton and Riichiro Mizoguchi (eds.), *Formal Ontology in Information Systems. Proceedings of the Sixth International Conference*. pp.400-413.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I.H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*. 11, (1).

Heja ,G., Surjam, G. & Varga, P. (2008). Ontological Analysis of SNOMED CT. *BMC Medical Informatics and Decision Makin.*. 8(8).

Keleberda, I., Lesna, N., Makovetskiy, S. & Terziyan, V. (2004). Personalized distanclearning based on multiagent ontological system. *Proceedings IEEE International Conference of Advanced Learning Technolog.*, pp.777-779.

Kabli, R., McCall, J., Herrmann, F. & Ong, E. (2008). Evolved Bayesian networks as a versatile alternative to partin tables for prostate cancer management. *Proceedings Genetic and Evolutionary Computation Conference.* pp.1547-1554.

Landis, J.R.; & Koch, G.G. (1977). "The measurement of observer agreement for categorical data". Biometrics 33 (1): 159–174.

Lukka, K. (2003). The constructive research approach. In L. Ojala, & O.-P. Hilmola,

Case Study Research in Logistics Series B 1, pp. 83-101.

Meyfroidt, G., Guiza, F., Ramon, J. & Bruynooghe, M. (2009). Machine learning techniques to examine large patient databases. Bailliere's *Best Practice & Research Clinical Anaesthesiology*. 23(1), 127-143.

Middleton, S.E., Shadbolt, N.R. & DeRoure, D.C. (2004). Ontological user profiling in recommender systems. *ACM Trans,Information System.,* 2(1), 54-88.

Pal, D., Chakraborty, C. & Mandana, K., (2011). Data mining approach for coronary artery disease. *Proceedings International Conference on Image Information Processin.,* pp.1-6.

Palaniappan, S. & Awang, R. (2008). Intelligent heart disease prediction system using data mining techniuqes. *International Journal of Computer Science and Network Security*. 8, 343-350.

Rasmussen, C.E. & Williams, C.K.I. (2006). Gaussian processes for machine learning. MIT Press.

Sathyadevi, G. (2011). Application of cart algorithm in hepatitis disease diagnosis.

Sundaram, S. (2010). Application of CART algorithm in blood donors classification. *Journal of Computer Science.* 6(5), 548-552.

Simpson, C.R., Anandan, C., Fishbacher, C., Lefevre, K. & Sheikh, A. (2007). Will Systematized Nomenclature of Medicine-Clinical Terms improve our understanding of the disease burden posed by allergic disorders. *Clinical and Experimental Allergy*. 37, 1586–1593.

Tao, X., Li. & Y., Zhong, N. (2011). A Personalized Ontology Model for Web Information Gathering. *Knowledge and Data Engineering, IEEE Transactions*. 24 (4), 496-511.

Ulrich, K.T & Eppinger, S.D. (2004). *Product Design and Development,* third ed. Boston: McGraw-Hill/Irwin.

Wang, J., Li, M., Hu, Y., & Zhu, Y. (2009). Comparison of hospital charge prediction models for gastric cancer patients: neural network vs. decision tree models. *BMCHealth Services Research*. 9, 1-6.

World Health Organization (2008).

Yuefeng, L. & Zhong, N. (2006). Mining ontology for automatically acquiring Web user information needs. *Knowledge and Data Engineering, IEEE Transactions.* 18, (4), 554-568.

Zadeh, L.A. (2004). Web intelligence and world knowledge-the concept of Web IQ (WIQ). In *Proeedings Of North American Fuzzy Information Processing Society.* 1, pp.137-146.

Zhong, N., Liu, J., & Yao, Y.Y. (2002). In search of the wisdom web. *Computer*. 11(35) pp. 27-31.

Appendix A: Infectious Diseases vs. Number of Synonyms in SNOMED CT

Infectious Disease	Number of Synonyms in SNOMED CT
Cholera	7
Diptheria	1
H5N1 Influenza	0
Japanese Encephalitis	4
Leprosy	3
Malaria	2
Measles	2
Meningitis	17
Mumps	32
Pertussis	9
Plague	17
Poliomyelitis	36
Congenital rubella	5

syndrome	
Rubella	24
Neonatal Tetanus	0
Infectious Disease	**Number of Synonyms in SNOMED CT**
Total Tetanus	14
Tuberculosis	80
Jungle Yellow Fever	2

Appendix B: Number of Synonyms in SNOMED CT vs. Number of Occurrences of Disease

African Region

Number of Synonyms	Number of Occurrences of Disease
0	
1	n/a
2	60769115
3	29814
4	n/a
5	n/a
6	
7	160801
8	
9	19425
10	
11	
12	
13	
14	5428
15	
16	
17	82312
24	16297

Number of synonyms	Number of occurrences
32	n/a
36	841
83	595184

South East Asia Region

Number of synonyms	Number of occurrences
0	
1	
2	3068200
3	166155
4	
5	
6	
7	
8	
9	
10	

Number of synonyms	**Number of occurrences**
11	
12	
13	
14	
15	
16	
17	
24	
32	
36	49
83	2124371

Region of the Americas

Number of synonyms	Average number of occurrences
0	20
1	48
2	561737
3	40474
4	
5	20
6	
7	
8	
9	7338
10	
11	
12	
13	
14	559
15	
16	
17	
24	18

Number of synonyms	Number of occurrences
32	31384
36	
83	201543

European Region

Number of Synonyms	Number of Occurrences
0	
1	41
2	7500
3	
4	
5	17
6	
7	
8	
9	29229
10	
11	
12	
13	

Number of synonyms	Number of occurrences
14	181
15	
16	
17	
24	11623
32	41448
36	
83	328796

Eastern Mediterranean Region

Number of Synonyms	Number of Occurrences
0	
1	109
2	7563872
3	4029
4	
5	
6	
7	

Number of synonyms	**Number of occurrences**
8	
9	9790
10	
11	
12	
13	
14	1194
15	
16	
17	
24	2030
32	
36	
83	411137

Western Pacific Region

Number of Synonyms	Number of Occurrences
0	1,792
1	129
2	1718968
3	5212
4	4220
5	
6	
7	2739
8	
9	40669
10	
11	
12	
13	
14	
15	
16	
17	

Number of synonyms	Number of occurrences
24	73077
32	413230
36	
83	1350929

Appendix C: WEKA Ready .arff file

```
@relation 'diseases'

@attribute date

{2007 2008 2009 2010}

@attribute region

{1 2 3 4 5 6}

@attribute synonyms real

@attribute occurrence real


@data

2009 1 0 2

2009 1 3 7

2009 1 3 4

2009 1 7 5

2009 1 9 4

2009 1 14 2

2010 1 17 4

2009 1 24 4

2010 1 36 1

2009 1 83 6

2009 2 0 1

2009 2 1 1

2009 2 3 4
```

2009	2	5	1
2009	2	9	3
2009	2	14	1
2009	2	24	1
2009	2	32	4
2009	2	83	5
2009	3	3	5
2010	3	36	1
2009	3	83	6
2009	4	1	1
2009	4	5	1
2009	4	9	4
2009	4	14	1
2009	4	24	4
2009	4	32	4
2009	4	83	5
2009	5	1	1
2009	5	3	2
2009	5	9	3
2009	5	14	2
2009	5	24	2
2009	5	83	5
2009	6	0	2

2009	6	1	1
2009	6	3	3
2009	6	7	2
2009	6	9	4
2009	6	24	4
2009	6	32	5
2009	6	83	6
2008	1	0	2
2008	1	4	7
2008	1	3	4
2008	1	7	5
2008	1	9	4
2008	1	14	3
2009	1	17	4
2008	1	24	4
2009	1	36	1
2008	1	83	5
2008	2	0	1
2008	2	1	1
2008	2	5	5
2008	2	3	4
2008	2	5	1
2008	2	9	4

2008	2	14	1
2008	2	24	2
2008	2	32	4
2008	2	83	5
2008	3	0	2
2008	3	1	3
2008	3	3	0
2008	3	3	5
2008	3	4	2
2008	3	7	2
2008	3	9	4
2008	3	14	3
2009	3	36	1
2008	3	83	6
2008	4	0	1
2008	4	1	1
2008	4	3	3
2008	4	4	1
2008	4	5	1
2008	4	9	4
2008	4	14	1
2008	4	24	4
2008	4	32	4

2008	4	83	5
2008	5	0	2
2008	5	1	1
2008	5	3	6
2008	5	3	2
2008	5	9	3
2008	5	14	2
2008	5	24	2
2009	5	36	1
2008	5	83	5
2009	6	0	2
2008	6	1	1
2009	6	9	6
2008	6	3	3
2008	6	4	2
2008	6	7	2
2008	6	9	4
2008	6	14	2
2008	6	24	5
2008	6	32	5
2008	6	83	5
2007	1	0	2
2007	1	9	4

2007	1	3	4
2007	1	7	5
2007	1	9	4
2007	1	14	3
2008	1	17	4
2007	1	24	2
2007	1	36	1
2007	1	83	5
2007	2	0	2
2007	2	3	1
2007	2	3	2
2007	2	5	1
2007	2	9	4
2007	2	14	1
2007	2	24	4
2007	2	32	4
2007	2	36	1
2007	2	83	5
2007	3	1	2
2007	3	3	6
2007	3	3	5
2007	3	4	2
2007	3	7	2

2007	3	9	4
2007	3	14	3
2007	3	36	1
2007	3	83	5
2007	4	0	1
2007	4	1	1
2007	4	1	3
2007	4	5	1
2007	4	9	4
2007	4	14	1
2007	4	24	4
2007	4	32	4
2007	4	36	1
2007	4	83	5
2007	5	0	1
2007	5	1	1
2007	5	4	6
2007	5	3	2
2007	5	9	4
2007	5	14	2
2007	5	24	4
2007	5	36	1
2007	5	83	5

2008	6	0	2
2007	6	1	1
2007	6	3	3
2007	6	4	2
2007	6	7	2
2007	6	14	2
2007	6	24	4
2007	6	32	5
2007	6	36	1
2007	6	83	5

Appendix D: Weka ready Regions

Region	Value
African region	1
Region of Americas	2
European region	3
South east Asia	4
Eastern Mediterranean region	5
Western pacific region	6

Appendix E: Range of Occurrences

Range of occurrences	Value
0-999	1
1000-4999	2
5000-9999	3
10000-99999	4
100000-999999	5
1000000-9999999	6
10000000-70000000	7