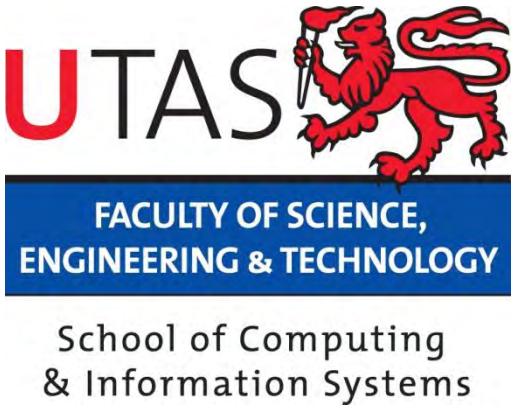


A Method for Knowledge Discovery and Development with Health Data

A dissertation submitted to the Faculty of Science, Engineering and Technology, University of Tasmania in fulfilment of the requirements for the Degree of Doctor of Philosophy.

Tristan Ronald Ling

BComp (Hons, First Class)

October 2011

Declaration of Originality and Access Authority

This thesis contains no material which has been accepted for a degree or diploma by the University or any other institution, except by way of background information and duly acknowledged in the thesis, and to the best of the my knowledge and belief no material previously published or written by another person except where due acknowledgement is made in the text of the thesis, nor does the thesis contain any material that infringes copyright.

This thesis may be made available for loan and limited copying in accordance with the Copyright Act 1968.

Tristan Ronald Ling

October 2011

Statement of Ethical Conduct

The research associated with this thesis abides by the international and Australian codes on human and animal experimentation, the guidelines by the Australian Government's Office of the Gene Technology Regulator and the rulings of the Safety, Ethics and Institutional Biosafety Committees of the University.

Tristan Ronald Ling

October 2011

Abstract

One of the most overlooked problems in the field of knowledge discovery is the acquisition and incorporation of existing knowledge about the data being analysed (Fayyad, Piatetsky-Shapiro et al. 1996; Pohle 2003; Kotsifakos, Marketos et al. 2008; Marinica and Guillet 2009). Doing this efficiently and effectively can greatly improve the relevance and usefulness of the results discovered, particularly for complex domains with a large amount of existing knowledge (Adejuwon & Mosavi, 2010; C. Zhang, Yu, & Bell, 2009). This study applies the successful Multiple Classification Ripple Down Rules (MCRDR) knowledge acquisition method to build a knowledge base from a complex dataset of lung function data, and describes a method for utilising the dataset to provide additional knowledge validation. The method acquired knowledge successfully, but indicated that a focus on rule-driven knowledge acquisition may adversely affect the MCRDR process. Knowledge acquisition was performed with multiple domain experts, with separate knowledge bases successfully consolidated using an evidence-based method to quantify differences and resolve conflicts. This knowledge comparison method was also tested as a learning and assessment tool for a small group of medical students, with positive results. In addition, the consolidated expert knowledge base was applied to the analysis of the lung function data, with a set of common data mining techniques, to reproduce and expand on a group of published lung function studies. Results showed that new knowledge could be discovered effectively and efficiently in a complex domain, despite the user having little domain knowledge themselves. Results were supported by recent literature, and include findings that may be of interest in the respiratory field. Notably, newly discovered knowledge is automatically incorporated into the knowledge base, allowing incremental knowledge discovery and easy application of those discoveries.

Acknowledgements

This work would not have been possible without the support of a huge number of people. Thanks must go to the staff of the Royal Hobart Hospital, Austin Health, and the TAHS study, for allowing access to their databases and for seeing potential in the analysis. Special thanks go to the respiratory specialists at the Royal Hobart for always being open, helpful, and willing to educate.

Special thanks also go to Stephen Giugni and the members of the CSIRO's Tasmanian ICT Centre, without whose support, interest, provoking discussions, and funding, this project would never have come as far as it has.

The researchers of the Menzies Research Institute and the University of Tasmania's School of Medicine also deserve enormous thanks. Every one of the (very many) people that were consulted in some manner about the project were helpful, friendly, and supportive, even in the face of endless requests, buggy software, terrible presentations, and even when they could not spare any time.

My supervisory team have been exceptional, and deserve exceptional acknowledgement. I simply could not have produced this thesis without them. To Dr Byeong Ho Kang: I sincerely thank you for the opportunities you have afforded me, for providing motivation when it was needed, and most of all for having faith in me. Without your support in particular, I would not have this thesis. To Dr David Johns: enormous thanks for your enthusiasm and support, even when I know at many times I was making no sense at all. Whenever I ran into problems that seemed to me insurmountable, you always came through and had everything running smoothly. I cannot imagine how I could have done this without you. To Dr Justin Walls: you astound me with your work ethic and capacity to rapidly assess a situation and find a solution. I have learned an incredible amount from you during our meetings, and you have never failed to identify and resolve the most important issues (often before I even understood them). You have provided me with impeccable advice, support, guidance, and frankly, something to aspire to. Thank you very much.

For the advice and support he has given me, massive thanks must also go to Dr Ivan Bindoff. For everything from technical discussions, motivation, amusement and

hilarity, and much, much more besides, you have helped me through like no one else, and I thank you hugely for that.

I would also like to acknowledge the role that my many and varied friends have played in the progress of this project. I particularly would like to thank Pops, for always knowing what I'm thinking, Princess for reading my thesis even without any prawns in it, Adam for being the most enthusiastic person I have ever met (and the best coconut player going around), Kingus for being Kingus, and Jamie for following a football team that we can always beat. To the nerds of clan FGI (and drnuk too I suppose): Dabe, Manny, Duncan, Mick, Anthony, and Smurf; and to the card players at lunch: Matt, Rob, Joel, and to everyone else that is not already mentioned (sorry), thank you very much for making my life as incredibly enjoyable as it has been. I have had the most amazing amount of fun, and it is all because of you.

To my family, I offer more thanks than I can give. You have all supported me in everything I have done, and I appreciate that more than I could ever tell you. Mum and dad, thank you so much. Noggin and Boo, Tanya and Richard: thank you, I really hope to see you all soon.

Finally to my Claire Bear, I cannot express how grateful I am for your support. I absolutely, without question, could not have finished this thesis without you. When I started this project, I was doing it for myself, but by the end I was doing it for us. Thank you for everything. I cannot wait to have free time with you again.

Table of Contents

Figures

Tables

Chapter 1 Introduction

Data is currently being generated at an unprecedented and ever increasing rate (Hilbert & López, 2011): it seems that new electronic records are created about us every day. Much of this data is accumulated and archived for a variety of reasons; key amongst these is the hope that an analysis will reveal patterns, which can aid in future decision-making (Witten & Frank, 2005; D. Zhang, Zhou, & Nunamaker Jr, 2002). Unfortunately the data is often of such volume that analysis is difficult and time consuming (Dai, Yang, Wu, & Hung, 2008).

Computational methods have been developed to assist in this analysis, under the headings of *knowledge discovery* and *data mining*. Knowledge discovery is the computer-aided process of finding new knowledge by analysing a set of data (Goebel & Gruenwald, 1999), and has seen a growth of popularity and development parallel to the growth of computing technology since the 1970s (Frawley, Piatetsky-Shapiro, & Matheus, 1992; Tukey, 1977). Data mining methods are computational algorithms which extract patterns from sets of data (Witten & Frank, 2005), representing the core of a knowledge discovery method; however it has long been identified that this is only one component in a much larger process, including components such as preparing the data for analysis, and interpreting the mined patterns to identify new knowledge (Fayyad, Piatetsky-Shapiro, & Smyth, 1996b).

The first step in a knowledge discovery process is to develop an understanding of the domain, which involves the identification and encoding of any relevant existing knowledge about the data. While this can greatly improve the effectiveness of the knowledge discovery, it is a step which has often been overlooked (Fayyad, et al., 1996b; Piatetsky-Shapiro, 1990; Pohle, 2003). Methods have been developed to apply existing knowledge to the computational analysis of data (Kotsifakos, Marketos, & Theodoridis, 2008; Liu, Hsu, & Chen, 1997; Marinica & Guillet, 2009; Piatetsky-Shapiro & Matheus, 1994), but the acquisition of this knowledge is a difficult and time consuming process, making knowledge discovery impractical for many complex domains (Adejuwon & Mosavi, 2010; C. Zhang, et al., 2009). The problem remains that knowledge discovery methods are largely ineffective in complex domains, as they lack the ability to acquire and incorporate the requisite domain knowledge.

A significant contributor to the difficulty of acquiring existing knowledge is that people with detailed expertise have very limited availability, given a typically high demand to apply their knowledge in practical situations. The efficiency of acquiring knowledge can be improved by taking input from multiple experts; however, this presents the potential for conflicts between different experts' opinions. Resolving these conflicts correctly and to both groups' satisfaction may not only beneficial to the development of a strong base of knowledge for data analysis, but may also be beneficial for improving the knowledge of the experts involved.

Once acquired, computer encoded knowledge can also be applied in a variety of ways. Expert systems are computer systems that can reproduce human expertise and apply it to complex tasks (Bobrow, Sanjay, & Stefik, 1986; B. G. Buchanan, et al., 1983; Luconi, Malone, Morton, & Michael, 1984). In this study, the method of acquiring expert knowledge is not tied to the knowledge discovery process, allowing the application of the acquired knowledge base as an expert system to assist in the interpretation of complex data.

The medical field presents unique challenges and benefits for knowledge discovery (Cios & Moore, 2002a). Archives of medical data are continually being added to, as the analysis of this data can provide solutions to the singularly important problems of life and death (Cios & Moore, 2002a; Roddick, Fule, & Graco, 2003). For these reasons the application of knowledge discovery methods are particularly relevant. However, analysing medical data is difficult as the data is complex, including a large number of measurements of a variety of types. Analysing the data is particularly difficult as extensive existing knowledge is needed to make meaningful interpretations of the data (Cios & Kacprzyk, 2001; Cios & Moore, 2002a; Prather, et al., 1997).

1.1 Thesis Structure

This thesis describes the development and testing of a method for discovering new knowledge for complex domains. To address the issue that existing knowledge needs to be incorporated into the process, the presented method involves acquiring, comparing, and consolidating the knowledge of multiple domain experts to develop a reliable knowledge base. This expert knowledge is acquired through Multiple Classification Ripple Down Rules (MCRDR), a common and effective knowledge

acquisition method (Kang, 1996; Richards, 2009), with some enhancements to provide additional data-based validation. A new method is also presented for quantifiably comparing multiple MCRDR knowledge bases and assisting in conflict resolution. The acquired and consolidated knowledge base is applied in numerous ways: as an expert system; to the discovery of new knowledge from a large compiled dataset; and as a teaching and assessment comparison for acquired student knowledge. These components and their interaction are described in Figure 1-1. The methods have been tested in the medical field of lung function, through the use of a compilation of archived databases.

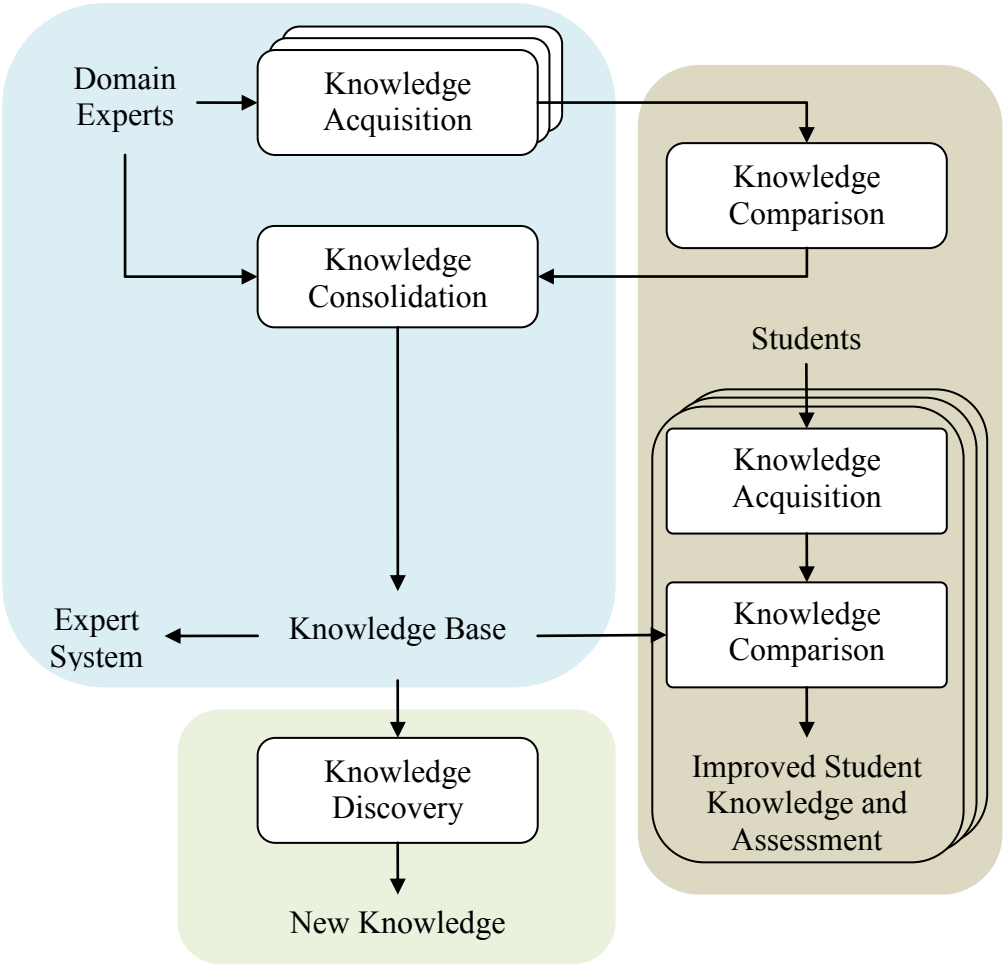


Figure 1-1: The structure of the methods presented in this thesis

Chapter 2 reviews relevant literature in knowledge discovery, knowledge acquisition, and the application of these methods to the field of lung function and the wider medical domain. Chapter 3, described by the upper-left blue segment in Figure 1-1, presents the process of knowledge acquisition and knowledge

consolidation used, and evaluates the effectiveness of the method in developing an expert system for lung function interpretation. Chapter 4, described by the lower green segment in the figure, discusses a method for applying this knowledge base to the analysis of a large database of lung function data and evaluates the effectiveness of the method in deriving new lung function knowledge. Chapter 5, described by the tan segment to the right of Figure 1-1, presents the method used to quantifiably compare the knowledge of multiple experts, and evaluates the efficacy of the method in identifying and assisting in the resolution of knowledge conflicts. The same chapter also describes how that knowledge comparison method was applied to identify the differences between the acquired knowledge bases of a group of medical students, and the expert knowledge base; and a discussion is presented on the application of the method as a learning and assessment tool. Finally, Chapter 6 summarises the findings of each component and discusses potential future developments and applications.

Chapter 2 Literature Review

Discovering new knowledge from data is a complex process involving many stages. This study presents an approach to knowledge discovery that focuses on the initial stages of identification and incorporation of existing knowledge. As such, this chapter will be devoted to an analysis of the previous research in the area of knowledge discovery, with a focus on knowledge acquisition methods, defining both the goals of the various fields of research and what methods are available to achieve them. As much of the work carried out in this study is focused on the study of lung function, a section is also devoted to explaining relevant research in that area, and the details that make this field a complex, interesting, and potentially valuable one to study.

2.1 Knowledge Discovery

The stated goal of finding new knowledge from data falls under the blanket term of *knowledge discovery*. In the most general terms, the field of knowledge discovery describes methods for finding new information about a subject, using some combination of recorded data on the subject and knowledge about that data (Goebel & Gruenwald, 1999). Ostensibly, the field therefore includes any method that can derive new data; but with the vast amount of archived electronic data available for analysis, and constantly being added to, the term is almost exclusively used to denote research into the development of computerised methods for data analysis and logical inference. These computerised methods are the focus for the majority of this section.

There are various terms used to summarise the field, each with its own connotations: knowledge discovery, *knowledge discovery from databases* (KDD), and more recently *knowledge discovery and data mining* (KDDM) are approximately synonymous; for a more in-depth examination of the differences between these terms, the reader is referred to Kurgan and Musilek's paper (Kurgan & Musilek, 2006). This study will use the term *knowledge discovery* for simplicity, although the intended meaning is perhaps closer to KDDM as described by Kurgan and Musilek.

The term *knowledge* is generally used in this study to refer to a relationship between data elements that represents some feature of the domain: be it the definition that

the value of one data element is exactly double that of another; a less specific expectation that, for a wide distribution of data, one data element will have a negative correlation with another; that the concept represented by one data element is a sub-concept of the concept represented by another data element; or any other similar relationship. Fayyad et al qualified this definition by saying that they considered a data pattern to be knowledge if it exceeded an interestingness threshold, adding that their definition is ―by no means an attempt to define knowledge in the philosophical or even the popular view", and is purely a practical definition for finding effective results (Fayyad, Piatetsky-Shapiro, & Smyth, 1996a). Nevertheless, this definition seems to have been largely accepted in the knowledge discovery literature (Fayyad, et al., 1996b; McGarry, 2005; Pohle, 2003; Prather, et al., 1997; Stumme, Wille, & Wille, 1998). While there are many other definitions of knowledge (Cassam, 2009; Ortega y Gasset & García-Gómez, 2002; Pears, 1971; Piaget, 1972; Zagzebski, 1999), we will restrict ourselves to the definition described here, with the qualification of Fayyad et al, as it is relevant to the knowledge acquisition and discovery tasks. Following this definition, computational knowledge discovery is the process of analysing data to find new, useful information about the topic that data is describing. Knowledge is considered new if it was not previously known to the person analysing the results of the knowledge discovery. Information is generally defined as any data that has some meaning, and therefore usefulness. Again, this implies the inclusion of a person who is interpreting the data. For philosophical discussions on these points see the aforementioned treatises; but the role of a person in knowing information and interpreting data will be discussed further throughout this work.

2.1.1 History and Context

Initial work in computational knowledge discovery focused on applying computing power to assist existing methods, such as making statistical analysis faster and more reliable and visualising data (Tukey, 1977). Knowledge discovery methods improved after adopting technology from expert systems development, whereby knowledge is acquired from an expert and applied to computational tasks (Frawley, et al., 1992). The 1990s saw a very rapid period of development and implementation of computing technologies in a huge range of fields, particularly in business fields such as marketing, manufacturing, and investing (Fayyad, et al.,

1996a), and an enormous increase in the quality of data recording and storage (Hilbert & López, 2011). With strong business successes for simple data analysis tools, there was also a significant rise in the quantity of data being stored: with most enterprises hopeful that an analysis of their business data could reveal ways for them to improve their outputs, processes, and income (Goebel & Gruenwald, 1999). With this explosion of data, there was a similar increase in knowledge discovery research; particularly into methods to analyse and derive information from large databases (Fayyad, et al., 1996a; Frawley, et al., 1992).

Within knowledge discovery, those methods which focused on database analysis became known as Knowledge Discovery in Databases (KDD) methods (Frawley, et al., 1992), a name first established by the first KDD workshop in 1989 (Piatetsky-Shapiro, 1990). The name was chosen to indicate that their goal was to take large amounts of data and discover knowledge (Fayyad, et al., 1996a), not just more data as in the case of data mining methods (which will be discussed in section 2.3.1, later in this chapter).

By the late nineties, research and development of KDD methods was a significant and growing field. Goebel and Gruenwald conducted a study of 43 different KDD software products, far from a complete list, and stated that ―despite its rapid growth, KDD is still an emerging field" (Goebel & Gruenwald, 1999).

KDD methods, in fact most knowledge discovery methods, follow a standard pattern for considering data. The data is broken into *cases*, typically consisting of a single transaction, event, or entity under study. The different pieces of information that make up the case are called its *attributes* (Witten & Frank, 2005). For example, a set of cases which represent books may have the attributes title, author, and price (although there will usually be many more attributes). Each case is defined by the values it has for the set of attributes – continuing the example, a particular case (book) may have the values: title: ―A Tale of Two Cities"; author: ―Charles Dickens"; price: $21.00. From here onwards, any use of the words case and attribute will likely be referring to these conventions.

2.1.2 Components of Knowledge Discovery

There have been many models developed to describe the knowledge discovery process, most of them quite similar. In a seminal work in 1996, Fayyad et al

presented the first of these, a 9-step, iterative model for effective knowledge discovery. In brief, the defined stages were: developing an understanding of the domain and the goal of the knowledge discovery; gathering a dataset; cleaning and pre-processing the data; selecting a subset of the attributes to be examined; matching the data and goals to suitable data mining methods; selecting a data mining method to use; running the data mining method with the data; interpreting the mined patterns to discover the identified knowledge; and consolidating and acting upon the discovered knowledge (Fayyad, et al., 1996b). This process was partially summarised in the diagram presented in Figure 2-1. It is clear from this that the data mining stage is only one step in the overall process and that all other steps are required for the method to be successful (Fayyad, et al., 1996b).

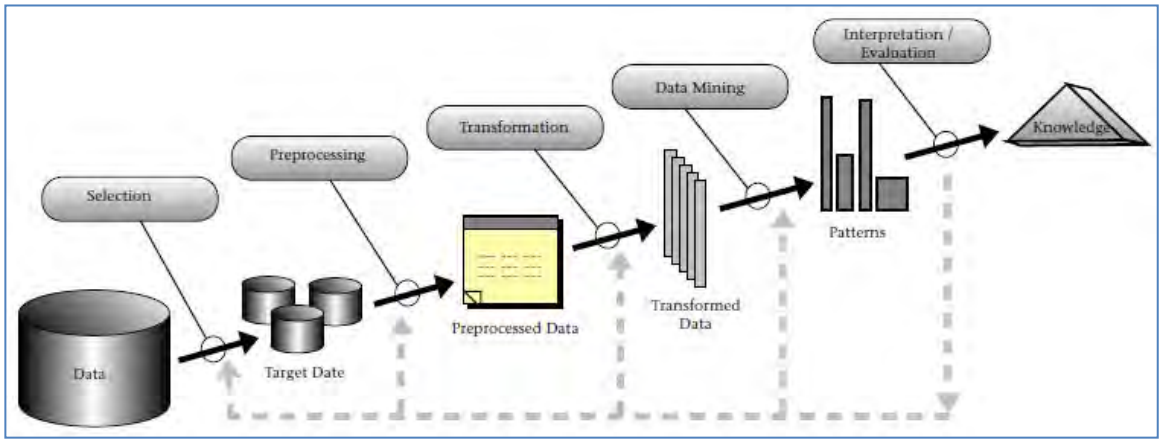


Figure 2-1: The Steps of KDD (Fayyad, et al., 1996a)

All subsequent models could be said to conform to the general structure described by Fayyad et al, although there are numerous distinctions of varying magnitude. In 2006 Kurgan and Musilek performed a detailed examination and comparison of the various models available, including those developed for both industry and research purposes, and presented a generic model based on the models considered. The 6 steps detailed in this generic model were (Kurgan & Musilek, 2006):

- Understanding the domain
- Understanding the data
- Preparing the data and selecting a data mining method
- Performing the data mining
- Evaluating the results

- Consolidating the discovered knowledge and applying it

A comparison between this model and the model of Fayyad et al finds relatively minor differences, primarily only in the emphasis and segmentation of different components: in particular, the generic model incorporates steps 3-6 into a single step. This thesis utilises the steps of the generic model when describing the process of knowledge discovery, although with different terminology at times.

2.1.2.1 Understanding the Domain and Data

The first steps in the process are some of the most vital, as the existing knowledge that is identified about the domain has a significant impact on every other stage: the selection, pre-processing, and cleaning of the data can be improved by having a better understanding of how the attributes relate; the data mining process can be improved as identifying the relationships that already exist can provide additional data to analyse; the interpretation of results is easier as new patterns can be related to existing patterns and presented in more understandable terms; and finally the establishment and understanding of existing knowledge can be used to determine how the new knowledge can be integrated, which areas it affects and how it might be used. These benefits are, after all, why this is the first step of the process. Despite this usefulness however, it has often been noted this step is often overlooked (Fayyad, et al., 1996b; Piatetsky-Shapiro, 1990; Pohle, 2003).

Many studies have identified that one of the most difficult issues in knowledge discovery is that far too many potentially interesting results are generated by statistical or mathematical data analysis; and in order to reduce this to a quantity that can realistically be evaluated, existing domain knowledge must be used (Liu, et al., 1997; Matheus, Chan, & Piatetsky-Shapiro, 2002; Piatetsky-Shapiro, Matheus, Smyth, & Uthurusamy, 1994; Piatetsky-Shapiro & Matheus, 1994; Pohle, 2003; Silberschatz & Tuzhilin, 1996; Sinha & Zhao, 2008). This would seem to be a logical conclusion: in order to determine which relationships are new, the existing relationships must be identified; and in order to determine which information is useful, knowledge about how that information might be used is required. Another component of this is that the stated goal of many studies is identifying ~~"in~~interestingness" (Freitas, 1999; McGarry, 2005; Ohsaki, Sato, Kitaguchi, Yokoi, & Yamaguchi, 2004; Piatetsky-Shapiro & Matheus, 1994; Tan & Kumar, 2001),

which is itself a subjective term dependent on a human expert (Freitas, 1999; Liu, et al., 1997; Pohle, 2003).

Approaches to Including Knowledge

Despite the identified importance of domain knowledge, few studies have expanded on how it can be successfully incorporated (Sinha & Zhao, 2008). This section will examine those methods that have been developed.

In a 1994 study, Piatetsky-Shapiro examined a method for using domain knowledge to identify which results were actionable, and what the impact of such an action might be. This allowed the list of potentially interesting results to be minimised by evaluating which results might lead to a useful outcome. The knowledge was incorporated in the form of rules added by a domain expert, identifying relevant attribute values, and a level of deviation from an expected value, culminating in a descriptive conclusion and a probability of success to allow a ranking of usefulness; an example is provided in Figure 2-2 (Piatetsky-Shapiro & Matheus, 1994).

if measure = payments_per_case

sector = surgical_admission

measure value increased by more than 10%

then recommend:

A study is suggested for discretionary and high-cost surgery.

success probability: 0.4

Figure 2-2: A sample rule, translated into pseudo-code, demonstrating incorporation of domain knowledge to identify useful knowledge (Piatetsky-Shapiro & Matheus, 1994)

In 1996 Liu et al proposed a method whereby a user can define *General Impressions* (GI) to describe non-specific knowledge about a domain: for example, "*Savings* > → *Yes*" would be used in a loan decision system to imply that having a large amount of savings would lead to being granted a loan, even though the user cannot put an exact number on how large an amount is required. This would then be used to evaluate the surprisingness or unexpectedness of a generated rule, by

lowering the interestingness of a rule if it conforms to an already supplied GI (Liu, et al., 1997).

More recently, some studies have been published examining the problem. In 2008 and 2009, Marinica et al described three types of knowledge used in data mining: domain knowledge, describing the data; user beliefs, relating to the user's expectations about knowledge that might be discovered; and a series of operators defining specific types of user expectation. Marinica et al proposed that domain knowledge be incorporated through ontologies, following Gruber's definition of ontology (Gruber, 1993). The user's expectations are defined by *rule schemas*, consisting of an association or an implication between ontology concepts, which essentially describe groups of potential rules. The user can then use a set of operators to define their specific expectations for each rule schema, describing whether any rules of that form should be ignored, or ascribed a certain level of interest. The approach has been tested with two simple ontologies and datasets, and the work is ongoing (Marinica & Guillet, 2009; Marinica, Guillet, & Briand, 2008).

Also in 2008, Kotsifakos et al proposed a similar approach for using ontologies. In this study, Kotsifakos et al assert that methods such as Liu's and Piatetsky-Shapiro's are impractical as the user is required to provide a set of expectations each time an analysis is to be performed. The method validates potential rules by checking that the attributes used in the rule are connected, based on the concept structures defined in the ontology; the rule is rejected if no such link is presented, or if the attribute concepts are too far away from the main ontology node. The study notes that this may reduce the potential of discovering truly new and unexpected knowledge. On a generated set of 25 rules, this method automatically selected as interesting the same 5 rules as an expert manually selected (Kotsifakos, et al., 2008).

One important factor that was identified by both Liu and Piatetsky-Shapiro, in concurrence with previous research into acquiring knowledge (Compton & Jansen, 1989), is that the acquired and required knowledge will change over time and in different contexts (Liu, et al., 1997; Piatetsky-Shapiro & Matheus, 1994). As noted by Kotsifakos, this means that a new knowledge acquisition process will be required for each data analysis. The identification of all relevant existing knowledge, especially of subjective knowledge, such as user expectations about the nature of the data in the specific area being analysed, constitute a significant knowledge

acquisition task. This is supported by recent studies which identified that most data mining techniques are not practical for many real domains because of the quantity of knowledge required (Adejuwon & Mosavi, 2010; C. Zhang, et al., 2009). It is suggested that the task of acquiring an expert's knowledge about a complex domain, and their expectations about the impact of every potentially interesting combination of attributes, presents a more costly task than has been typically indicated in the literature; and for complex domains, such as medical areas, an impractical solution. Literature regarding the difficulties with knowledge acquisition, and a description of various methods developed in that field, will be discussed in section 2.2.

2.1.2.2 Analysis of the Data

The analysis of the data, or the data mining component, is the crux of a knowledge discovery method. As such it is usually given the most attention of any of these components; although it has been estimated that the data mining step only constitutes 15% to 25% of the overall work (Brachman & Anand, 1996). This step contains some sort of process that takes in data, and returns information derived from or about that data. Typically it seeks to identify patterns in the data, which can be interpreted to identify relationships that represent the knowledge that the method is trying to discover (Fayyad, et al., 1996b). However, the methods for doing this tend to produce a large amount of extraneous information: without an effective assessment of these results to determine their meaning, usefulness and applicability, the results are often meaningless and unusable (Goebel & Gruenwald, 1999; Piatetsky-Shapiro, 2000). This step will be covered in more detail in section 2.3.

2.1.2.3 Interpreting Results and Applying Discoveries

Once the data has been analysed and interesting patterns identified, regardless of the particular method used, the results need to be interpreted to find what knowledge has actually been discovered. This step was identified by Fayyad et al as a critical component of knowledge discovery in 1996 (Fayyad, et al., 1996a), but is one that has received little focus since (Pohle, 2003). One of the reasons for this neglect is that it is one of the most difficult components of knowledge discovery, as it inevitably requires expertise and human analysis (Piatetsky-Shapiro, 2000; Pohle, 2003); indeed, in more complex areas the identification of relevant patterns is no longer of concern, but rather the interpretation of those results (Subramanian, et al.,

2005). A related problem is the incorporation of newly discovered knowledge into a knowledge base, or assisting the user in identifying how to apply the new knowledge (Pohle, 2003). The process of discovering new knowledge, and having this knowledge directly incorporated into the process of discovering new knowledge, has been identified as a major goal for knowledge discovery systems, but one that is an open problem (Matheus, et al., 2002). A 2006 study used *fuzzy ontologies* to discover patterns, where elements could be matched to multiple concepts, each to differing degrees of confidence. This approach was described as occasionally making the resultant patterns more understandable for the user, but did not resolve the issue (Escovar, Yaguinuma, & Biajiz, 2006). These steps are considered again briefly in section 2.4.

2.2 Knowledge Acquisition

Identifying and formalising existing knowledge is a vital step which can completely alter how effective the knowledge discovery process will be. Many studies have highlighted the importance of this stage, and often also mentioned the lack of effort applied in this area (Brachman & Anand, 1996; Fayyad, et al., 1996b; Piatetsky-Shapiro, 1990; Thearling, 1998). Effectively identifying what is already known can dramatically improve the end results by ensuring that the process is not simply rediscovering knowledge that is already known. Also, increasing the base of knowledge to search from can, depending on the method used, increase the sophistication of what can be gleaned (Fayyad, et al., 1996b; Ordonez, 2006). It is worth considering however that the usefulness of existing knowledge can be dependent on the type of knowledge that is being searched for and the extent of knowledge already known: if too much emphasis is placed on using existing knowledge, this can constrain the search and thus reduce the range of knowledge that can be discovered. At times, it is desirable to utilise knowledge discovery methods which are unconstrained by any existing patterns and conventional wisdom (Piatetsky-Shapiro, 1990).

The type of knowledge and manner in which it is collated and applied can vary greatly. As will be discussed, for many approaches it is considered enough to identify which cases to consider, or which attributes to examine most thoroughly. However, the process can also be as detailed as adding much more data to each case,

by identifying case groupings and sub-groupings; inferring further information; and identifying and correcting errors within the data.

The process of collecting existing knowledge almost invariably involves a human with expertise in the area, as it is ultimately human knowledge which the method is trying to extend. The task then becomes one of converting human knowledge into a format usable by the knowledge discovery process, typically some form of software. In computing, the field of *knowledge acquisition* is concerned with methods that do exactly this – convert human knowledge or expertise into a machine-usable, reproducible format (Liou, 1990).

2.2.1 History

Knowledge acquisition has been defined as the process of ~~"~~extracting, structuring, and organising knowledge from human experts so that the problem-solving expertise can be captured and transformed into a computer-readable form" (Liou, 1990). Another way to express this idea is that knowledge acquisition is the process of modelling human expertise about a subject.

The field arose from work towards building Artificial Intelligence (AI). In the 1950s and 1960s there was the beginning of an understanding of the power of computational systems and the benefits they could provide in automation of tasks. However, due to the high cost of developing new systems, there was much discussion on how a computer might be made that could learn new tasks easily; particularly systems that could learn independently, without human direction (Friedberg, 1958; McCarthy, Hayes, & SCIENCE., 1968; M Minsky, 1961). These studies laid the foundations for many modern knowledge acquisition, machine learning, and expert systems techniques.

2.2.1.1 Expert Systems

During the 1960s there began an increasing focus on the practical applications of AI studies, as much of AI development had become mired in philosophical questions and produced few practical computer systems. Rather than attempting to create fully adaptable, independently thinking AI systems, some of the lessons learned were used to enhance the sophistication and capabilities of systems by teaching them how to perform specific complex tasks (B. G. Buchanan, et al., 1983; B. Gaines & Boose,

1988; Lederberg, Feigenbaum, & CALIF., 1967). The first example of this is the development of the DENDRAL system (Lederberg, et al., 1967), widely regarded as the first example of an *expert system*: software which can reproduce human expertise for complex tasks (Bobrow, et al., 1986; B. G. Buchanan, et al., 1983; Luconi, et al., 1984). The tasks performed by these systems are typically something that most people cannot perform, one which requires special training or study: hence the word *expert*. Expert systems have also been defined as requiring problem-solving abilities, to distinguish them from *knowledge-based systems* (Davies & Darbyshire, 1984), which are described as any repository of knowledge. However, as the functional difference between the two often only depends on the interface used to access them, and the situations in which they are accessed, the two terms are often used synonymously. Many have further determined that an important feature of an expert system is that it has the ability to explain its reasoning: the theory behind this ultimately coming from Plato's definition of *knowing* (B. G. Buchanan, 1986; Davis, Buchanan, & Shortliffe, 1977).

Expert systems are occasionally intended to replace a human expert in performing a task, and while it is debatable that this is the ultimate goal of all expert systems research, it is much more common for them to be deployed to assist experts in their work; in which case they have the more specific name of an *expert support system* (ESS).

History of Expert Systems

The first expert systems were developed from the late 1960s as a way to apply computational power to help solve complex real problems. As previously mentioned, the earliest recognised expert system was DENDRAL (Dendritic Algorithm), a system for assisting organic chemists in interpretation problems: applying chemical knowledge to elucidate molecular structures (Lederberg, et al., 1967). This system consisted of a collection of knowledge about the domain, in the form of rules or heuristics (which will be discussed shortly); an engine to be able to apply these rules, with some internal logic to produce a result; and an interface to allow users to input the data on which it bases its decisions (B. G. Buchanan & Feigenbaum, 1978).

Given the success of the method in assisting the organic chemists, this approach was widely adopted and produced many more expert systems in the 1970s and 1980s, with significant literature regarding the process and how it could be defined and improved. In 1986 Buchanan published a list of expert systems in routine use or field testing, along with a bibliography of expert systems literature; some 46 systems and 374 publications respectively (B. G. Buchanan, 1986).

With the extensive research and development in the area, the components of the expert system model became better defined. The compiled knowledge became known as the knowledge base: a collection of heuristics or classification rules, built by a person known as the *knowledge engineer*. The knowledge engineer was someone with programming proficiency in the language and structure of the system, who could interview or observe the expert performing their task and translate the task into the rules that the system could use. A programmer was also required to construct the inference engine which could infer, based on given data and the knowledge base, what the response should be. The front-end became known as an *expert system shell*: an interface for users, often the experts themselves, to be able to input the data being examined and receive the system's responses.

However in the late 1980s expert systems rapidly lost their popularity, as the methods in use proved incapable of meeting expectations: the limited applicability, expense of development, frequent failure to meet the standards of the human experts, and the marketing hype combined to bring a rapid fall in investment (Gill, 1995). A report in the *Wall Street Journal* in 1990 indicated that the expert systems field was probably worth $600 million: but that some researchers had estimated that it would have reached $4 billion (Bulkeley, 1990). This dramatic fall was a result of a series of problems. The systems could only function in the very specific domain which they were designed for: adding further capabilities to the system would require extensive further knowledge engineering, additions to the expert system shell, and potentially modifications to the inference engine. The knowledge engineering process was also a common point of failure: the translation between what was observed or described into a rule set, by someone who is not themselves an expert in the field, is a difficult process fraught with potential problems. It was also a very slow and expensive process: the engineer would be required to engage the expert for a considerable time, and then spend more time translating what was

learned into something the system could use. This interaction between expert, knowledge engineer and knowledge base became known as the ―knowledge acquisition bottleneck", as it was considered the most expensive and difficult component of expert systems development (B. G. Buchanan, et al., 1983; Lenat, Prakash, & Shepherd, 1985).

2.2.2 Knowledge Acquisition Methods

With the recognition that knowledge acquisition was the most critical component in the development of expert systems, many approaches were explored. This led to a widely branching field of methods for knowledge acquisition and expert system development. This section will explore the major developments in knowledge acquisition technology and how they have been applied.

2.2.2.1 Classification Rules

As the first successful expert system, DENDRAL provided the standard template for expert systems until the 1980s. DENDRAL's approach to knowledge acquisition involved a programmer, later known as a knowledge engineer, interviewing experts and encoding their expertise as rules (B. G. Buchanan & Sutherland, 1968). These rules were described as *heuristic* rules by the developers to indicate that they are not absolute laws or complete definitions: rather they are guidelines or suggestions that, through inferencing, typically lead to correct results (B. G. Buchanan & Sutherland, 1968).

These heuristic rules are an example of what are more commonly known as classification rules; and they are one of the first, and in many ways the simplest, of the data modelling and knowledge acquisition techniques. The term *classification* refers to the grouping of data cases by some measure: all the cases in a group are said to have that classification, or belong to that class. The term classification is used in many of the methods described here in exactly the same way; it is a common way of using knowledge about a case to add extra information, which may be used to derive new information or to make deriving further information more efficient. A classification rule is a rule which defines which cases should belong to a given class, and why (Witten & Frank, 2005).

A classification rule consists of one or more conditions, and a classification or conclusion. The classification is typically a name or number which is used to denote a given group of cases; such as in the DENDRAL system, each classification is an interpretation, or label, of what the case represents. Each condition in the rule consists of some statement describing a type of case, usually in the form [attribute name] [operator] [value], for example: *Author* (attribute) *is* (operator) "*C. Dickens*" (value).

It should be noted that there are many other terms used to describe different implementations of rules, for example *production rules* or *inference rules*; however in practice the rules take the same form. The differentiation comes in their application: an *inference rule* is one which is used to derive new information, which can then be used to cause another rule to activate, and so on until no more rules activate: at which point some of the additional information provided by the rules is presented as the classification (Witten & Frank, 2005). Although the term classification rule is used here, it is intended to cover each of these different named rules, all of which follow a very similar pattern.

The most common source for classification rules are human experts. Rules are usually easy to create and understand, often being close to literal natural language statements, and as such were very popular early in knowledge acquisition research and are still commonly used today (Davis, et al., 1977; Stansfield, 2009). The expert examines the dataset and creates rules which classify cases based on the values of the set attributes. For example, all cases with a sufficient value for attribute *A,* and where attribute *B* is negative, should have conclusion *1* (`if A>30 AND B<0 then` *1*).

One of the major advantages of classification rule systems is that the structure of the knowledge learned is readable by the expert – if the expert wants to know why a classification was made, they can simply examine the conditions of the rule that "fired" (the rule that provided the final classification) (Clancey, 1984): which is a critical component in creating useable, verifiable expert systems (B. G. Buchanan, 1986; Davis, et al., 1977). It is generally easy to view the compiled knowledge and see what conclusions are being made, and based on which knowledge: hence providing a simple means to review effectiveness and progress.

Another major advantage for using these rules is that they represent discrete pieces of knowledge; which allows new knowledge to be added simply by adding more rules, rather than rewriting significant portions of code; and allows modification of existing rules (B. G. Buchanan & Sutherland, 1968; Davis, et al., 1977).

A criticism of rule-based systems was that they lacked flexibility in their conclusions, and could not be applied to many real-world problems, especially reasoning problems, because facts and data are rarely certain (Zadeh, 1979). This led to further development by adding Bayesian probabilities (discussed in section 2.3.3) and *fuzzy logic* (in which a classification is provided a confidence, between 0 and 1, rather than simply being present or not present). Each method gave alternative means of adding likelihoods or certainty factors to knowledge and to classifications, greatly improving the applicability and usefulness of results; however this came at the cost of more complex inferencing and knowledge acquisition (Duda, Hart, & Nilsson, 1976; Zadeh, 1979).

As work progressed on implementing more complex rule-based systems, and as existing systems were added to, it also became apparent that in order to solve complex tasks thousands of rules might be needed (Walser & McCormick, 1977). As it was realised that each rule condition and classification was typically reused many times, this led to strategies to reduce the storage requirements of the rules, such as the *inference net* (Duda, et al., 1976). These structures assisted in being able to process and use very large numbers of rules, but did not solve other aspects of the problem. The acquisition of these rules from human experts was very time-consuming and risky, as the translation between expert, knowledge engineer, and rule was not exact. Anything that was missed or entered incorrectly added to the increasingly difficult maintenance of a system where it became very difficult to predict what the effect of adding a new rule or changing an existing rule might be. It also became apparent that no matter how much effort was made to be thorough in the knowledge acquisition, the knowledge bases were never complete: due to changing knowledge, new discoveries, and fallible human memory, there would always be new rules that would have to be added (Davis, et al., 1977; Duda & Shortliffe, 1983). The maintenance of such a system therefore became an ongoing and very difficult task: each new rule that was added needed to be extensively checked and modified by a knowledge engineer to ensure that it would not change

the result of any other rule. This process was facilitated by the *cornerstone case* system, whereby any case that caused the inclusion of a new rule would be stored; whenever a new rule was to be tested it would have to be compared to every cornerstone case to ensure that it did not match and change the results for the previously reviewed cornerstone cases – an exhausting process (Compton & Jansen, 1989). As a result of these acquisition and maintenance difficulties, the knowledge acquisition component became accepted as being the bottleneck in expert systems development (B. G. Buchanan, et al., 1983; Lenat, et al., 1985; Walser & McCormick, 1977), and alternatives to rule-based system knowledge acquisition began to be explored.

2.2.2.2 Decision Trees

One of the first developed alternatives to rule-based systems was the decision tree, which became popular in the 1980s as a potential solution to the problems in expert system development. Using the decision tree method, a logical tree is formed consisting of nodes and branches: an attribute is associated with each node, and for each possible value (or range of values) for that attribute a branch is created leading to a lower node. The lowest nodes have no outward branches, and contain a classification (Carbonell, Michalski, & Mitchell, 1983). In this manner, a case can be presented to the tree, and by following the branches appropriate to the values for the case, a classification is found. Hence, knowledge is stored in a relatively simple to follow format, and one which can easily be transformed into a graphical representation such as in Figure 2-3 (Quinlan, 1986). These features provide the explanatory requirement of an expert system, and the incremental nature of the development allows the expert to have input into the way that knowledge is structured at each step (Quinlan, 1986).

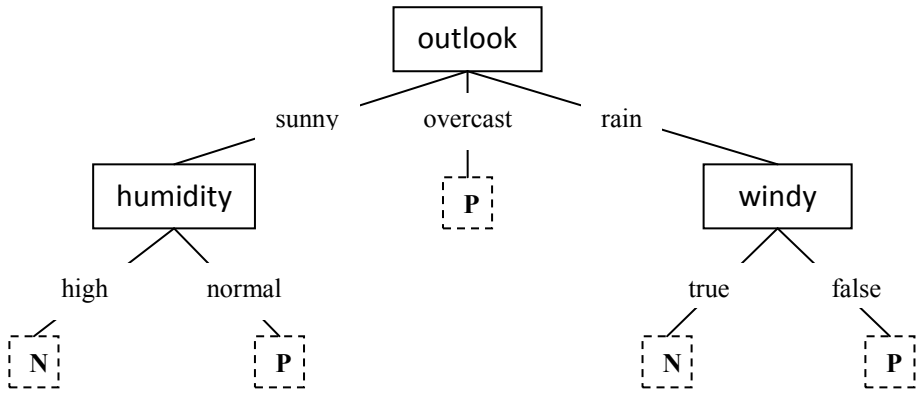


Figure 2-3: An example decision tree for choosing a positive (P) or negative (N) result based on weather conditions (Quinlan, 1986)

Decision trees had already existed before the 1980s; initially they had simply been considered a poor alternative to rule-based systems, due to the structure being more difficult to understand and define than classification rules, and more difficult to maintain (Hart & Center, 1977). For a time decision trees were still considered an unviable option, denounced by some as not worthy of being called an expert system (Hayward, 1985). However, given the knowledge acquisition problems faced by the alternative methods of the time, decision tree development was renewed; particularly after the creation of Quinlan's C4.5 algorithm for induction of decision trees (Quinlan, 1986).

C4.5 became a popular method because it helped ameliorate the knowledge acquisition bottleneck. Rather than trying to collate and store an expert's knowledge, with C4.5 the expert looked through a large set of data and classified it; then the method used statistical methods to derive a decision tree that was correct for all (or as much as possible) of the data. This had the potential to greatly reduce the time required to acquire expert knowledge, particularly if such a classified dataset were already available (Quinlan, 1986). However the acquired knowledge takes little advantage of the knowledge that the expert has, and is unlikely to accurately represent any of the decisions which the expert makes; which was the source of much of the criticism from rule-based systems researchers (Duda & Shortliffe, 1983; Hayward, 1985; Prerau, 1985). While the decision tree could still provide an explanation of why it chose a particular classification, such a response may not have any meaning to an expert, who used an entirely different approach to their task; and

for a task of realistic size and complexity, the decision tree can become very large and convoluted, such that a trace of which decisions led to the classification would be almost meaningless. Another drawback is that the inevitable maintenance is a difficult process (Quinlan, 1987; Witten & Frank, 2005). For these reasons, decision trees are an effective method to use for some problems, producing a system that can complete tasks effectively with minimal development expense; but in most cases are an ineffective knowledge acquisition tool.

2.2.2.3 Case Based Reasoning

Another knowledge acquisition approach developed at around the same time is case based reasoning (CBR). CBR shifted the focus on the components of knowledge from the role of inferencing to the role of memory, basing decisions on past experiences rather than incremental logical inference (Kolodner, Simpson, & Sycara-Cyranski, 1985). This approach was developed from work in cognitive sciences by Schank on the nature and role of memory (Schank, 1980). In CBR knowledge is represented by a set of stored cases and a set of defined classifications or solutions. In some domains, this approach has been found to be closer to the manner in which experts already consider their domain; and hence knowledge acquisition is an easier process for the expert (Kowalski, 1991).

The fundamental procedure for CBR methods is that each successfully solved or classified case is stored, along with how it was solved. Then, when any new case is examined, the system need only find the most similar case and apply the same solution. If necessary, the retrieved solution can be revised to fit this new case; and once solved, the case and its solution are stored (Kolodner, 1992).

As a knowledge acquisition approach, the expert examines cases individually and inputs each of them, one at a time, into the system. The CBR system compares the current case with all the previously stored cases in the knowledge base, and produces a list of cases that it considers to be similar to the new case. The method would then use the classification(s) of those cases as the classifications for the new case. The expert examines this, comparing it with their own opinion of what the classification should be for the current case. If the expert believes there is an error or something is missing from the system's logic, the expert corrects or adds this knowledge as appropriate. This is achieved by adding the current case to the set of

stored cases and changing which classifications apply to it (Kolodner, 1991). This process is shown in Figure 2-4.

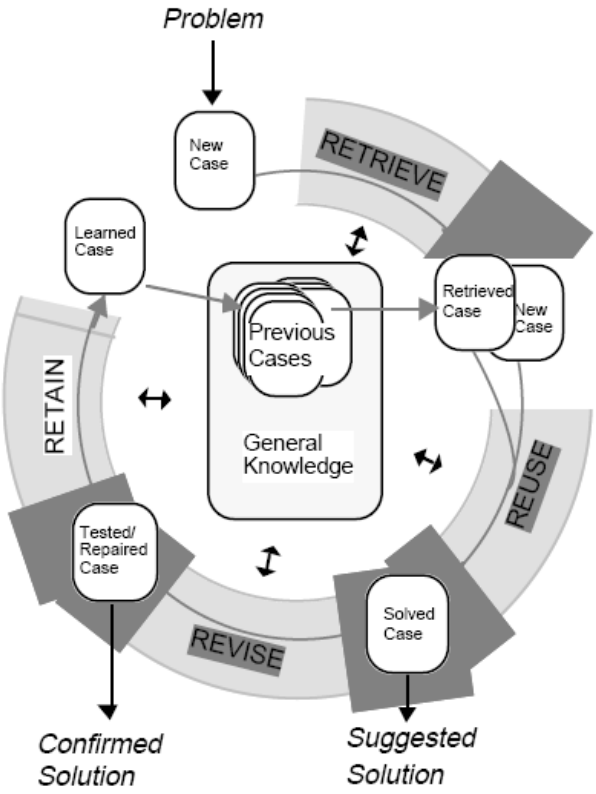


Figure 2-4: The Case Based Reasoning Process (Aamodt & Plaza, 1994)

This approach can be divided into two distinct categories: precedent-based CBR and problem-solving CBR (Rissland & Skalak, 1989). Precedent-based CBR focuses on identifying similar past cases and using them as evidence to justify using a given solution; whereas problem-solving CBR focuses on the solutions used and adapting the existing solutions to fit the new case (Barletta & Mark, 1988).

There are a few points of note about this approach. It can be seen that CBR is itself not a method: it contains no prescription on how to identify which cases are similar, leaving this up to the individual implementation. In this sense, CBR is more of a philosophy or general approach than an actual method. The retrieval stage is clearly the critical component, as this is where the expertise is applied: while the knowledge can be considered to be the cases and their solutions, this knowledge cannot be usefully applied to a new case without an effective means of identifying which other case it is most similar to. Hence, while CBR makes an effective knowledge acquisition method for associating cases with solutions, there still needs

to be a further knowledge acquisition step for the knowledge of what constitutes similar cases.

There is also some difficulty in easily influencing the results that the method provides (Féret & Glasgow, 1997; Golding & Rosenbloom, 1996; Kolodner, et al., 1985; Manago, et al., 1993; Yamaguti & Kurematsu, 1993). It should also be noted that the expert is necessarily involved in the process, to review the solution and revise it if necessary, which generally makes maintenance of adding to the knowledge base quite simple; although this again depends on how the system is implemented. Similarly, CBR meets the criteria of being able to explain its result, albeit in a slightly different manner to previous methods: the CBR system can show its reasoning by presenting the past cases that were used to generate that classification, and the attributes that were used to identify the two cases as most similar (Kolodner, 1992).

A drawback of CBR is that the knowledge base is entirely dependent on previously seen examples which may only represent a small subset of the dataset, rather than representing the entire domain (Chi & Kiang, 1991). While this can often be said of any technique using a dataset, it is particularly apparent in CBR due to the method being based entirely on the cases themselves: the knowledge for how to correctly resolve case types not yet seen cannot be entered into the system.

2.2.2.4 Ripple Down Rules

Similar ideas to Schank's cognitive science studies led to other developments in knowledge acquisition. At around the same time as CBR was being formalised, a method with similar attributes was being developed which significantly reduced the problems associated with the knowledge acquisition bottleneck.

Ripple down rules (RDR) were developed by Compton and Jansen in the late 1980s and early 1990s, after experiences in maintaining a rule-based expert system GARVAN-ES1 (Horn, Compton, Lazarus, & Quinlan, 1985). They observed that even though the system had a 96% accuracy rating when it was introduced, within 4 years the accuracy had been improved to 99.7% by the addition of extra rules by the knowledge engineers. The problem was that in order to achieve that 3.7% improvement, the number of rules in the knowledge base had doubled (Jansen & Compton, 1988). Compton and Jansen made the important observation that no rule

could be guaranteed to be correct, with most rules subject to revision given enough time in use; some rules were even removed entirely at a later point by different experts (Compton & Jansen, 1989). They proposed that the knowledge that experts provide during knowledge acquisition is rarely, if ever, a complete representation of their understanding, but rather just a justification for why their conclusion is correct, in the context of the current case (Compton & Jansen, 1989). This was an alternative explanation for the extensive maintenance that was always necessary with rule-based systems.

In order to resolve this problem they proposed a new method of structuring and maintaining rule-based systems, based on the idea that knowledge was context-dependent and would require maintenance. The rules in a RDR knowledge base are structured in a binary tree, in which each node contains one rule. RDR knowledge acquisition functioned in a similar manner to the maintenance of a typical rule-based system: the system would be presented with a case, which it would attempt to produce a classification for based on its knowledge base and provide its reasoning. The expert would examine the result, and if they disagreed with the system's conclusions, the system would be updated with a new rule. The difference is the manner in which the rule is added:

- If the system had no classification for the rule, then the expert would simply add a rule with the correct classification, justifying why the case should have that conclusion. This rule would be added in a node as the child to the right of the rightmost node in the tree.
- If the system reached an incorrect classification, the new knowledge is added in the context of the incorrect rule, as this is the situation in which the expert is being asked to justify their classification: the rule is added as the child to the left of the incorrect rule, as an *exception* to the incorrect rule.

Note that the choice of left and right are arbitrary, as long as they are consistently used: more correctly there is a *true* or *rule satisfied* direction, and a *false* or *rule not satisfied* direction. This process is described in Figure 2-5.

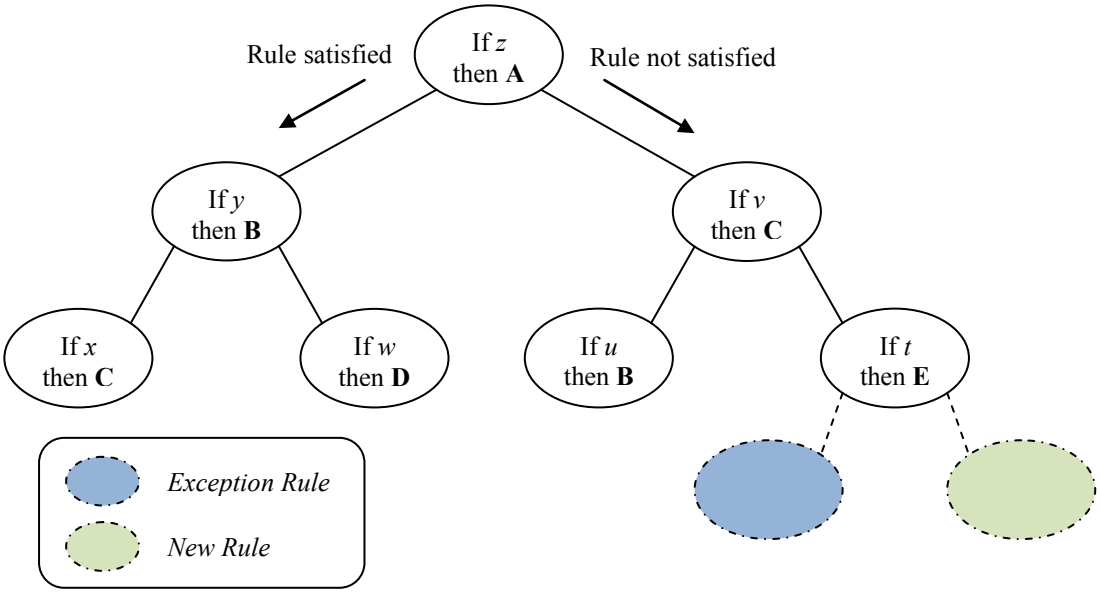


Figure 2-5: An RDR Knowledge Base. If classification E is reached incorrectly (i.e. if a case does not have attribute *z*, nor *v*, but has attribute *t*), then an *exception rule* will be added in the blue node. If no classification is reached, a new rule will be added in the green node.

When considering what results the knowledge base can determine about a case, the first rule considered is the top level of the tree; if it is satisfied, then the child rule on the left branch is considered. If the top level rule is not satisfied, the child on the right hand branch is considered. This continues until no further child nodes' rules are satisfied, or no further child nodes exist. The case is classified according to the classification associated with the node whose rule was last satisfied (Compton & Jansen, 1989). This structure ensures that any new rule will not interfere with any rule other than that it was added to correct: removing the necessity for a knowledge engineer to laboriously check that new rules do not break existing knowledge and allowing knowledge acquisition and maintenance to be carried out by the expert as the system is in use.

An alternative way of considering the structure is shown in Figure 2-6. In practice, the tree is typically heavily weighted on one side, as more new rules are typically required than corrections (Compton, et al., 1991). Because of this it might be easier to consider an RDR tree as a list of rules, each with a correction tree branching from them.

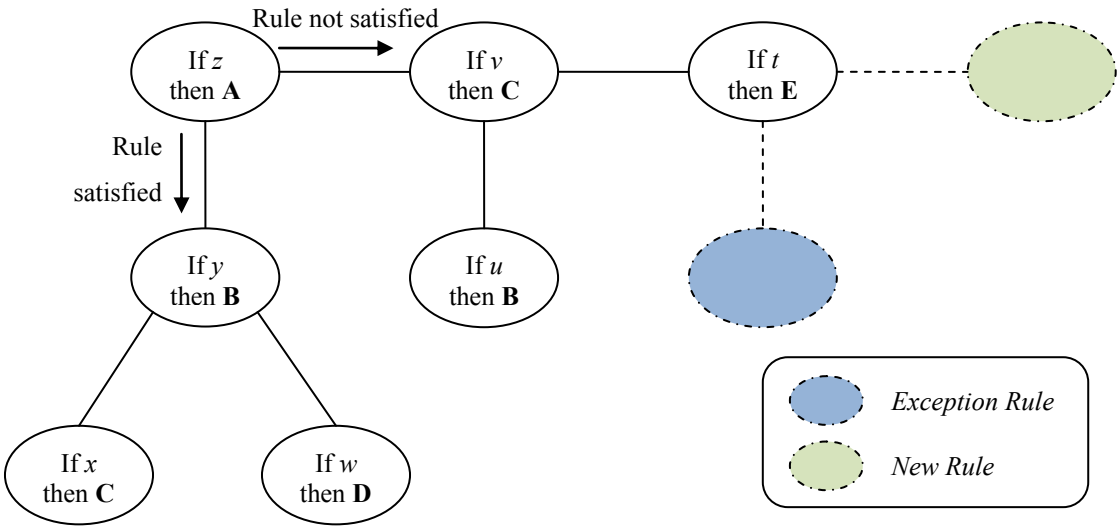


Figure 2-6: An RDR Knowledge Base presented as a list of rules with correction trees

A further advantage of this structure was that no knowledge engineering was required; the expert's justifications were simply added directly to the knowledge base as rules, without an engineer laboriously checking that existing rules were still valid. RDR achieved this by maintaining evidence for the knowledge base: the case used to define each rule was associated with that rule and called the *cornerstone case*. Whenever a rule was found to classify a case incorrectly, and the expert offered a justification for why it was wrong, that justification was compared against the cornerstone case for the system's rule. If the new rule changed the classification for that cornerstone case, this was indicated to the expert. The expert was then required to provide further details justifying the different classifications for the two cases, or to accept that the previously seen case was incorrectly classified.

This new method was applied to redeveloping the GARVAN-ES1 expert system, and it was found to require no knowledge engineering: the expert's rules were simply added directly, not requiring any validation. This resulted in ~~k~~knowledge acquisition at least 40 times as fast as that required for a conventional version of the same knowledge base, with the same knowledge engineer/expert involved" (Compton, et al., 1991). While an increase in efficiency would be expected, due to previous experience developing the same expert system, this is a very substantial improvement.

Although the knowledge base can be confusing to view in this structure, the system can still easily provide explanations for its classifications by tracing back the path through the tree and providing the rules at each node. Such rule traces are often better descriptors of the knowledge as the rules are exactly as entered by the expert, containing no extra conditions added for engineering reasons (Compton, et al., 1991). It was also noted that this approach automatically, if gradually, discovers tacit knowledge: knowledge that experts hold but which they find difficult to express, or are not consciously aware of outside of the specific context in which the knowledge applies (Richards & Busch, 2003).

PEIRS

Possibly the most well-known RDR health expert system, and perhaps the best known RDR system in any domain, is the Pathology Expert Interpretative Report System (PEIRS). PEIRS was an expert system for interpreting chemical pathology reports, and was an early and major success with the RDR methodology (Compton, et al., 1992; Kang, Compton, & Preston, 1995). It established a knowledge base of thousands of rules by having multiple experts interact with it directly over a period of a few years. It achieved an accuracy rate estimated to be greater than 95%, for over 20% of the pathology domain: an impressive feat making it one of the largest and most successful expert systems in routine use at the time (Compton & Edwards, 1994). This early success was later built on with commercial applications, particularly with further developments to the RDR method (Compton, Peters, Edwards, & Lavers, 2006).

Drawbacks

Although the knowledge base built easily and maintenance was simple, building a complete knowledge base still took considerable time (Richards & Compton, 1997a). Another problem is that some classifications would likely have to be entered multiple times, if the same classification occured as an exception to multiple rules (Compton, et al., 1991; Kang, et al., 1995). One of the most significant drawbacks is that, because of the rule structure, only one classification could be reached for any one case. For most domains, this caused the creation of many complex rules, each implying a compound classification: although the knowledge base may contain rules which can classify the case in multiple ways, it

will always use the first classification that it reaches (Compton, Kang, Preston, & Mulholland, 1993). This related to a further problem: the knowledge base did not allow true inferencing, as each rule would only be considered once (if at all), with only the first classification reached being provided as the result (Compton, et al., 1993).

RDR Modifications

As RDR represented a significant improvement in knowledge acquisition and maintenance, the method was applied and modified in many ways, each overcoming some flaw. Gaines developed a method for automatically inducting a ripple down rules knowledge base, in the same way as C4.5 could induct a decision tree (B. Gaines & Compton, 1992) – this will be discussed further in section 2.3.2. The lack of true inferencing in the RDR method led to a number of modifications, to allow the method to be applied to configuration tasks. Recursive RDR used several (8) individual RDR knowledge bases, each contributing a part of the final configuration solution. The results of these knowledge bases could also be used as data for the other knowledge bases, allowing a kind of inferencing (Mulholland, Preston, Sammut, Hibbert, & Compton, 1993).

Nested RDR (NRDR) provided a similar solution, by allowing the expert to define a separate RDR knowledge base for each concept. This concept knowledge base would then decide whether that concept was applicable to the current case or not. Thus each concept could be used as a condition in rules defining other concepts: as the RDR knowledge base produces a single classification, the value for this classification can be used as a rule condition (Beydoun & Hoffmann, 2000). Both Recursive RDR and Nested RDR provided a solution to the problems raised by configuration problems, but had their own drawbacks (Bindoff, 2010).

A further addition to the RDR method was an attempt to give the knowledge acquisition system some understanding of its own limitations, later described as the Prudent RDR method (Compton, Preston, Edwards, & Kang, 1996). This method sought to resolve a more general knowledge acquisition issue: that expert systems lack common sense and cannot tell when their classifications are obviously wrong and they should be asking for more information. After reaching a classification for the current case, the prudent method identified whether the current case was a

typical example of the cases that usually provide that classification: if the new case has a value for an attribute that is not present in the cases that have been confirmed with the classification, the expert is flagged that the classifications may be wrong and should be checked. Successful tests led to a conclusion that this is a useful aid in knowledge acquisition (Compton, et al., 1996).

Multiple Classification Ripple Down Rules

The most significant problem with RDR was that it could only provide a single classification; this was particularly evident as the domains in which it had first been applied were both multiple classification problems (Compton, et al., 1993; Kang, et al., 1995). This was a topic of much discussion in the early 1990s, and soon an adaptation of the RDR technique was developed, named multiple classification ripple down rules (MCRDR) (Compton, et al., 1991; Kang & Compton, 1992; Kang, et al., 1995).

The problem was resolved by generalising the binary tree structure to an n-tree. When a classification is found it is recorded, but the search through the tree is continued. In this structure, all children nodes represent exceptions rather than just the leftmost node, and new (independent) rules are added as siblings at the top level of the tree.

The method's approach is to consider the rule for the current node: if its rule is satisfied, then all its children nodes are considered, in turn. If their rules are satisfied, then their children nodes are considered, and so forth. Any rule which has its own rule satisfied, but not those of any of its children, has its classification added to the result list. This is shown in Figure 2-7Figure 2-7: A MCRDR KBS. The highlighted boxes represent the rules that are satisfied for the case [a,c,d,e,f,h,k]. The final classifications are classes 2, 5 and 6 (Kang, et al., 1995). Another way to consider this is that running a case through the knowledge base is a depth-first search, where child-links are only followed if the child's rule is satisfied; and any rules which are satisfied, but have no satisfied children, have their classification added to the result list.

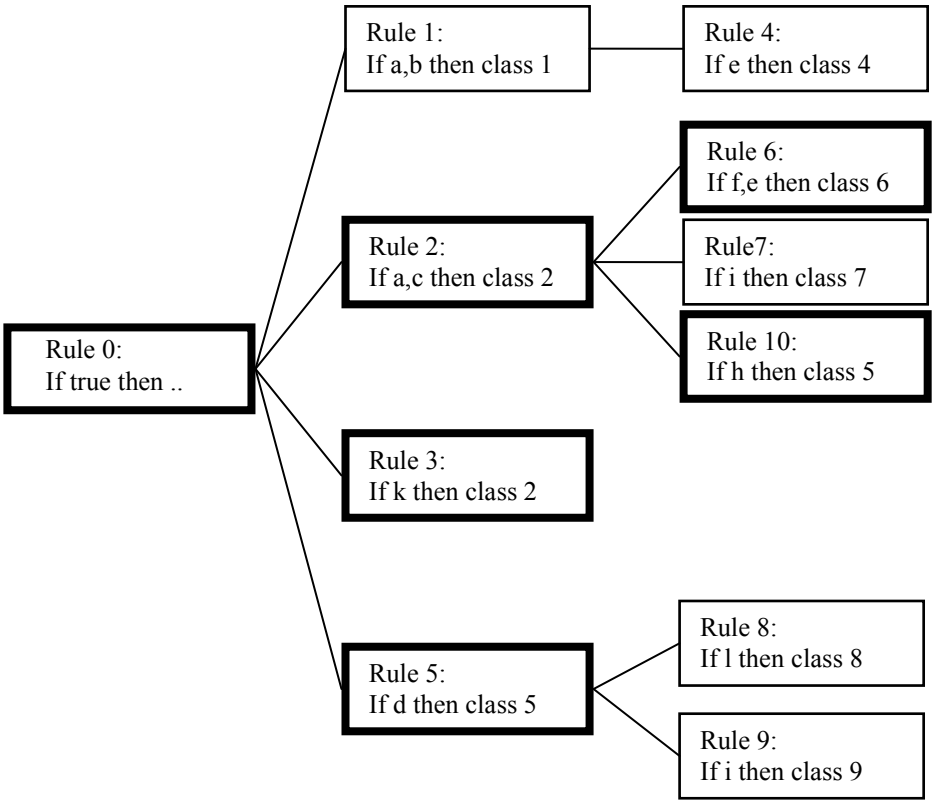


Figure 2-7: A MCRDR KBS. The highlighted boxes represent the rules that are satisfied for the case [a,c,d,e,f,h,k]. The final classifications are classes 2, 5 and 6 (Kang, et al., 1995)

This structure maintains the acquisition of knowledge in context; but certain changes needed to be made to the knowledge acquisition process to support them. Adding a new classification to a case is straightforward: the rule is added at the top level of the tree, as a sibling to the other top-level rules. Correcting a rule is also simple, being quite similar to single classification RDR (SCRDR): the new rule is added as a child to the incorrect rule (i.e. in the context that it is required); and as with SCRDR, the rule only needs to be sufficient to differentiate a case from the parent rule as it will only be considered within that context. Allowing multiple classifications also presented a third option during knowledge acquisition: a classification may be wrong or extraneous, requiring no classification to replace it but nevertheless being incorrect. In this situation the expert defines what was named a *stopping rule*. This is identical to an exception rule in all ways, except that it has no classification. The rule is added in the context of the incorrect or extraneous classification, and its conditions are defined by the expert as an explanation of why the given classification should not appear for the current case. Table 2-1 below

summarises these three situations, and the appropriate action to take (Kang, et al., 1995).

Wrong classifications	To correct the KB
Wrong classification to be stopped	Add a rule (stopping rule) as an exception to the incorrect rule, to prevent the classification
Wrong classification replaced by new classification	Add a rule as an exception to the incorrect rule, to give the new classification
A new independent classification	Add a rule at the root level to give the new classification

Table 2-1: The three situations in which new rules can be added to a knowledge base (Kang, et al., 1995)

The allowance of multiple classifications also impacted the ensured validity of an RDR knowledge base. In single classification RDR any new rule only needs to be checked against the cornerstone for the rule it is correcting, and new rules do not need to be checked at all, as the rule will only ever be applied within the context of the other rule firing or no rules firing respectively. As MCRDR does not stop checking rules after a single classification however, the addition of a new rule can potentially affect the results for many cornerstone cases. To resolve this, the new rule needs to be checked against any cornerstone case which could reach the context in which the rule is added: in other words, all cases which satisfy all of the antecedents of the new rule. If any cornerstones which match all antecedent rules also match the new rule, this must be brought to the expert's attention as it indicates a knowledge acquisition error: either the new rule is lacking some knowledge, and needs to be refined so it can differentiate between the cornerstone cases' classifications and the new classification; or the cornerstone case was originally misclassified and should in fact have the new classification.

In SCRDR each cornerstone is linked to the rule that is added for it, greatly improving the efficiency of validating against cornerstones as only the cornerstone which must be checked is checked. In MCRDR, each cornerstone can satisfy

multiple existing rules when it is reviewed, and multiple rules that are added after it has been accepted. In order to maintain an efficient review of cornerstones, it is clear that a cornerstone case must not only be linked to the rules that were defined from it, but all rules which it currently satisfies: both existing rules and rules that are added later. Thus, when a new rule is added, all cornerstone cases which satisfy the new rule must also be linked with the new rule; and all existing rules which the new case satisfies, and their antecedents, must be linked to the new case. Links to a cornerstone case are not deleted when that cornerstone satisfies an exception to that rule, as any further exceptions to the rule may also apply to the case (Kang, et al., 1995).

Both the SCRDR and MCRDR methodologies have some distinctive traits that differentiate them from other knowledge acquisition approaches. They have many similarities to case based reasoning approaches: cases are considered sequentially and individually, and as much knowledge as possible is extracted from a case before moving onto the next (Kolodner, 1992; Kowalski, 1991). The CBR-like philosophy is continued with the strongly evidence-based approach to validating knowledge, as implemented by the cornerstone case system: rather than attempting to validate new rules against existing rules, new rules are validated against previously seen cases. If the classifications for a previously seen case are modified by a new rule, conditions must be added to the new rule to differentiate the previous case from the current case, or the previous case must be accepted as having been misclassified (Kang, 1996). This improves the completeness of new knowledge and ensures consistency with existing knowledge, but without requiring any knowledge engineering. Another distinctive trait is that once knowledge is entered into a knowledge base, it is never deleted. As the knowledge structure is built with exceptions, the outcomes of the knowledge base can always be changed by adding new rules. Therefore as long as the context of knowledge is maintained correctly, there is no need to delete anything, and no risk that legitimate knowledge will be removed (Compton & Edwards, 1994; Kang, et al., 1995).

There remains a concern with MCRDR knowledge bases that knowledge may need to be added repeatedly, as exceptions in different locations or as groups of conditions used multiple times. It has been found however by multiple studies that MCRDR knowledge bases are typically very compact and efficient structures: in a

comparison with a SCRDR knowledge base for a single classification task, it was found that the MCRDR knowledge base was smaller (Kang, 1996); and a study was undertaken in the late 1990s to automatically compact and reduce knowledge bases to resolve the issue, but it was found that even after extensive work the improvements were fairly minimal, and for most situations not worth the effort (Suryanto, Richards, & Compton, 2002).

Modifications

As it represented a significant step forward in applicability with few drawbacks the MCRDR method was widely adopted, seeing commercial application in a number of domains, such as chemical pathology. Compton's 2006 paper details the integration of a commercial RDR system into one laboratory (of many who adopted the product), where it was routinely used to assist in the interpretation of the laboratory reports. This single instance of a commercial application saw some 16,000 rules and 6,000,000 cases over a 29 month period (Compton, et al., 2006). Although, as with most research developments, there are generally few details of the commercial applications, the success of MCRDR still saw a number of variations developed which sought to overcome perceived shortcomings. MCRDR was adapted to a resource allocation task with the Sisyphus-I system (Richards & Compton, 1999), which led to much discussion about how MCRDR could be applied to configuration tasks (Compton, et al., 1998; Ramadan, et al., 1998). It was also adapted to a help-desk system, under the name Interactive Recursive RDR, integrating Recursive RDR and adding the ability to ask the user questions to derive further information (Vazey & Richards, 2005). Further developments are discussed below.

Rated MCRDR

Rated (or Weighted) MCRDR was developed in the early 2000s as a way of identifying relationships between classifications, in the form of a value, with the initial application of identifying how important a case is to the expert (Richard Dazeley & Kang, 2003). The example used was an email classification system which could identify advertising spam emails or important work-related emails: if this system received work-related advertising it would receive both classifications; the new method would be able to give a rating of importance to the emails, ranking

work emails highest, advertising lowest, and work-related advertising between the two. The method used to find this value was to build a perceptron neural network (discussed further in section 2.3.3), taking a weighted value from each rule in the knowledge base as the inputs.

This theory was extended to include discovering different kinds of new information in later studies, particularly to developing a Prudent RDR-like estimation of whether a case represents a situation not covered by the current knowledge base (R Dazeley & Kang, 2004). This approach was shown to perform slightly better than the Prudent-RDR system, but not well enough that an expert could only examine the flagged cases.

FCA

Formal concept analysis (FCA) is a mathematical method for identifying and displaying concepts. It was developed in the early 1980s by Wille (Wille, 1982), who later applied it to knowledge acquisition tasks (Wille, 1989).

The FCA method is based on the philosophy of a concept consisting of the relationship between objects (*extension*), and the attributes belonging to those objects (*intension*). The extension is the set of objects which provide examples of the concept, and the intension is the attributes which define the concept. These definitions are used to develop *concept lattices*: graphs that display the relationships between the extension and the intension thereby revealing the concepts (knowledge) present in the domain.

This method was applied to acquired MCRDR knowledge bases to assist in the reuse of the knowledge outside of the constraints of the expert system the knowledge was acquired for (Richards & Compton, 1997b). The acquired MCRDR knowledge bases were converted into flat knowledge bases of rules, converting exceptions into individual rules. Sections of this were then chosen to be examined further, as the complete knowledge base was too large to be visualised.

The concept matrix produced used rules as extents and rule conditions as intents. The current concept matrix could be defined in two possible ways: the selection of a classification, or the selection of a rule. Selecting a classification first would select all rules reaching that classification as the extents, and each of the rule conditions of those rules as the intents. Selecting a rule would initially select each of that rule's

conditions as the intents, then search through the knowledge base for all occurrences of each of those conditions, adding the rules that contained them to the list of extents. The definition of this concept matrix then allowed the creation of a concept lattice, which was used to provide a better visual representation of the knowledge contained in the knowledge base (Richards, 1998), and to show the expert how the concepts in a new rule fit with existing concepts, for additional validation (Richards & Compton, 1997b).

Ontology Development

Since RDR was developed, there has been an increasing realisation that the knowledge acquired does not provide exact definitions, and will only be correct some of the time (when it is in the appropriate context). This led to the realisation that knowledge, as represented in a knowledge base, changes over time, and initial knowledge is frequently wrong. These insights caused the rejection of the conventional wisdom that knowledge acquisition required an initial phase of modelling the domain, or the development of an ontology, in order to be successful (Richards & Compton, 1997c; Suryanto & Compton, 2001). As the development of an ontology for a domain is still a desirable outcome, and the FCA studies showed that domain knowledge could be modelled from existing knowledge bases, studies were undertaken to develop ontologies from RDR knowledge bases after the knowledge acquisition, with some success (Suryanto & Compton, 2000, 2001).

Case And Rule-Driven (CARD) Systems

In 2006, Vazey examined the case-driven paradigm of MCRDR, and presented a predictive model for the case-driven acquisition of knowledge (Vazey, 2006). The model quantified the problem of striving for a complete knowledge base when acquiring knowledge from a stream of random cases. The study concluded that to be most effective, knowledge acquisition should also incorporate some rule-driven elements, creating a Case And Rule-Driven (CARD) approach.

In the same year, Vazey collaborated with Richards to develop a prototype help-desk knowledge acquisition and expert system based on the CARD model, with collaborative editing of a central knowledge base. Minimal results were presented, but it was reported that users were willing and able to use both the case-based and rule-based knowledge acquisition schemes (Vazey & Richards, 2006). An

examination of the collaborative component of the study is presented shortly in section 2.2.3

MCRRR

A recent development of the MCRDR method at the time of writing is a study by Bindoff (Bindoff, 2010), which added the capability of true inferencing to MCRDR. Titled "Multiple Classification Ripple Round Rules" (MCRRR), this addition allowed the classifications that a rule reaches to also be used as conditions for rules: the result being a more intuitive way of defining rules, and enhanced capabilities in configuration tasks; but at a cost of higher computation time. This approach shows promise in providing a more robust and general solution to the configuration tasks presented in Recursive RDR, Nested RDR and the like (Bindoff, 2010).

ProcessNet

Another recent movement has been to develop more complex, multi-tiered intelligent systems. Such intelligent systems generally consist of many different processes, each with their own knowledge acquisition requirements. In a recent study, the ProcessNet system was built to allow not only incremental improvement in the knowledge for each component process, but to also allow incremental learning for how those processes interact. The combination of processes is described by a directed, acyclic graph; by following the flow of cornerstone cases through these processes, and ensuring that each subsequent process is updated accordingly when a process changes, the ProcessNet approach provided a medical imaging analysis system that showed continued improvement even allowing for a large number of incremental updates to individual processes (Misra, Sowmya, & Compton, 2011).

MCRDR with Multiple Experts

RDR systems have been developed using multiple sources of expertise wherever possible; for large scale production systems it can be a necessity as no single expert can devote sufficient time to a project to develop a sufficiently complete and useful knowledge base. Most commonly however, expert availability is the biggest hindrance to the system development. As such, including multiple experts is often not a viable option. The standard approach has been to try and ensure that only the most knowledgeable experts use the system, but this requires a balance between

level of expertise and availability of expertise. Despite the difficulties, the inclusion of multiple experts is still desirable, as reducing the workload of the expert and accessing different points of view and experiences provide the potential for a much more effective expert system (Shaw & Woodward, 1988).

Traditional Approach

Many developed RDR systems have used multiple experts without any specific modifications to the process, including initial systems such as PEIRS (Compton & Edwards, 1994). The acquisition was performed as normal: whoever was currently using the system entered rules for the current case, regardless of whether their knowledge was exactly correct or agreed upon by other experts. If the knowledge was incorrect in some way it was expected to be caught by the in-built validation process, either as the rule was entered, or later when the rule incorrectly classified a case and was noticed by an expert who recognises the error. This strategy has obvious flaws. Firstly, it offers no guarantee that differences between expert opinions will be identified. More significantly, it offers no protection against a less knowledgeable expert changing a more knowledgeable expert's input to match their own view. Similarly, if incorrect knowledge is added but then corrected by a more knowledgeable expert, there is again no protection from the less knowledgeable expert reversing the correction back to the original error, as they view this to be correct. In the past it has perhaps been assumed that, as the experts are working with the same system in the same physical location, they are likely to have other occasion for communication when the knowledgeable expert can raise the issue and inform the other of the correct rule; or, if equally knowledgeable experts are disagreeing on a point, they will meet and discuss it to find a consensus. However, the method itself made no provision for informing an expert that their rule had been modified; presumably the assumption being that all errors would be discovered and corrected given sufficient time, making this just another component in the asymptotic approach of knowledge base accuracy to full correctness.

Knowledge Base Integration

Recently more work has been undertaken into the development of knowledge bases with multiple experts. In 2005 Beydoun and Hoffmann presented work into how to incorporate multiple experts' conceptual knowledge into a central ontology or model. The study used Nested single classification RDR (NRDR) knowledge bases

in a hierarchical concept structure (labelled Multiple Hierarchical Restricted Domains, or MHRD) to represent individual experts' knowledge. These models were then automatically integrated, where possible, through a method based on identifying and quantifying probable inconsistencies between those models. The probable inconsistencies were calculated by comparing definitions against all possible values for all attributes. Synonymous concepts were merged if the estimated differences between their definitions did not exceed a given threshold. If the difference did exceed the threshold, an expert was consulted about whether the concepts should be merged. The estimated degree of internal inconsistency was then used as a measure of the completeness and correctness of the knowledge in the integrated model (Beydoun, et al., 2005).

Collaborative Knowledge Acquisition

Also in 2005, Richards and Vazey began examining better methods for the collaborative development of knowledge bases. Their studies used Web 2.0 models such as *wikis* and social networking paradigms to allow multiple experts to work with a single central knowledge base (Vazey & Richards, 2005; Vazey & Richards, 2006). This collaborative approach was applied to an ICT support call centre. Numerous call centre staff were able to access a central knowledge base to find solutions for ICT support cases, and to then update the knowledge base if necessary. Any element of the knowledge base or the cases could be edited or removed. Conflicts were identified in two ways: by tracking all changes to cases or rules, and allowing a user to see the history of the changes made to any given element; and by allowing users to mark knowledge base elements as *live* or *registered*, with *live* elements being viewable to the public and *registered* only privately viewable, essentially indicating the status of an element as agreed upon or in conflict. Under the wiki/Web 2.0 paradigm, these conflicts were brought to the attention of the authors and left to be resolved through online discussion and other users' input. Any user could elect to receive notifications of changes to a knowledge base, element via email or SMS (Richards, 2009). However the focus of Vazey and Richards' work was not on ensuring that the users' knowledge was correct, but rather the end point of having a complete, current and correct knowledge base. This is partially a factor of the domain of development; within the ICT support call centre, individuals were not necessarily expected to have the knowledge required to solve a case, but rather,

to be able to find the relevant knowledge in the knowledge base (Vazey & Richards, 2006).

A further contribution of Vazey and Richards' work was the testing of the Case and Rule-Driven (CARD) paradigm of MCRDR development. The CARD methodology identifies that knowledge can be added via a top-down, rule-driven approach, or via a bottom-up case-driven approach. The case-driven approach is the normal paradigm of MCRDR development, but Vazey identifies that there are benefits to rule-based development and concludes that a combination of the two is likely to provide the best results, although no study has been performed comparing the two approaches (Vazey, 2006; Vazey & Richards, 2006). Collaborative knowledge base development is discussed again in Chapter 5.

2.2.3 Knowledge Comparison

The strong research and development focus on expert systems in the 1980s, and the identification of the difficulties involved in knowledge acquisition, also gave rise to research into the comparison and consolidation of knowledge. As there were a wide variety of knowledge modelling techniques available, this research generally focused on how to perform knowledge comparisons for each of those specific modelling techniques.

In 1989 Shaw and Gaines identified that when acquiring knowledge from multiple experts, each may describe different parts of their knowledge, use different terminology, or use terminology differently (Shaw & Gaines, 1989). They described four possible situations in acquiring knowledge from multiple experts: *consensus*, when the experts use the same terminology for the same concept; *conflict*, when experts use the same terminology for different concepts; *correspondence*, the use of different terminology for the same concepts; and *contrast*, the use of different terminology and different concepts. This scheme was applied to the knowledge of a group of experts, acquired as repertory grids: a technique which allows the definition of conceptual models by asking an expert to list what they considered to be the entities in their domain, then being asked to define distinctions between them (Fransella, Bell, & Bannister, 1979; B. R. Gaines, 1987; Shaw & Gaines, 1989). They concluded that any comparison of expert knowledge necessarily involves approximation, as evaluating a complete conceptual system is impractical: there

must be some level of assumption about underlying concepts, which may not in fact be identical. However, identifying significant similarities or differences is a valuable task as it promotes directed, contextual discussion among the experts that may reveal other more subtle distinctions (Shaw & Gaines, 1989).

Dieng in 1997 described a method for combining multiple experts' knowledge when that knowledge is represented as conceptual graphs (Dieng, 1997). Conceptual graphs are a technique for visually representing knowledge: at the simplest level, by defining concepts as graph nodes and relationships as the links between them, but conceptual graphs can also represent first order logic, and contain rules as reasoning. A concept graph contains a set of concepts, a set of relationships, and a set of *individual markers*, which indicate when a concept is a named entity rather than a type of entity (Chein & Mugnier, 2008). Dieng's study describes a detailed algorithm for how to combine multiple concept graphs, including comparing the concept set, relation set, and individual markers in turn, and identifying and resolving synonyms and homonyms in the names of the components (Dieng, 1997). A problem with concept graphs however is that they are difficult to develop, requiring significant work by a knowledge engineer in interviewing experts and attempting to elicit the conceptual models that the experts use.

Richards and Compton's combination of Formal Concept Analysis (FCA) and Ripple Down Rules (RDR), also in 1997, could also be used to compare the concepts in different experts' knowledge. The derived concept lattices of two knowledge bases could provide a visual representation of the concepts implied by each expert's rules, allowing easier visual identification of their differences (Richards & Compton, 1997c). This method was shown to be effective in the identification of broad conceptual differences, for example when an expert defines classes which another did not (Richards & Compton, 1997c). However this method is less effective at identifying subtler differences between expert's knowledge, and presents no information about the significance of each difference. For example, if two experts' knowledge bases displayed a minor difference in the values used in certain rule conditions, say one expert used $x<20\%$ and the other $x<25\%$, this could visually appear equally as significant a difference as one expert having an entirely new rule. The viewer also receives no information on the significance of these

differences: two knowledge bases may contain rules which use many subtly different conditions, yet almost invariably present identical results in practice. This is of course not a downfall in all circumstances: if examining how experts conceptually regard problems, the identification of those differences might present a significant result in itself. However, when comparing knowledge bases with a large number of differences, information on the significance of each difference may be needed to perform the comparisons efficiently and effectively.

Similarly, Beydoun and Hoffmann's method for automatically integrating multiple knowledge bases is applicable as a knowledge comparison method (Beydoun, et al., 2005). However, while this method worked well for automatically combining knowledge bases (as much as is practical), it made no provision for resolving conflicts or improving expert knowledge. Making comparisons using all possible values for all attributes is also a concern, as the maximum ranges of attributes are not always obvious and modelling them could take considerable effort. Also, without considering the likely distribution of values for each attribute, the resultant comparison may misrepresent the significance of a difference: a small difference in value for one rule condition may conceivably result in 100 cases classified differently or none, depending on where within the distribution the condition's value lies.

At approximately the same time, Vazey and Richards conducted studies into the application of the *wiki* paradigm to knowledge acquisition, whereby many experts can collaboratively update a central store of knowledge. In their approach, all parts of the knowledge base or cases could be edited or removed, with conflicts identified by tracking a history of these changes, or by marking the rules as accepted or in conflict. Identified conflicts were brought to the attention of the users who created the conflict, and resolved through online discussion and other users' input (Richards, 2009; Richards & Vazey, 2005; Vazey & Richards, 2006).

The ICT support domain however presented some quite different features to other application domains of MCRDR. The most fundamental difference is that each of the users have relatively little knowledge specific to each ICT problem, with their expertise focused primarily on general problem solving skills: a survey of users indicated that for 67% of cases the user would not have the knowledge to resolve the case, and would need to refer to other sources (Richards & Vazey, 2005). The

focus of the development therefore was to incorporate these other sources into the central knowledge base, making the knowledge base the primary source of knowledge. Thus, the most important goal of knowledge acquisition in the ICT support domain is to make the knowledge base as complete and correct as possible, without particular concern for the users' knowledge, as it is assumed that they will be retrieving their knowledge from the knowledge base. This subtly contrasts with the goal in other applications of MCRDR, such as the medical domain considered in this study, where the goal is to support an expert's knowledge and decisions rather than present authoritative solutions (Musen, Shahar, & Shortliffe, 2006). A further difference in the domain is in the outcome of a case. In ICT support, a case is correctly resolved once the problem is corrected. Unless the problem subsequently recurs, the solution can be said to be correct regardless of what the solution may have been. This does not always apply in other domains however. In a medical interpretation setting, the resolution of a case is often ambiguous: different experts may well provide different interpretations, and there is often no conclusive evidence as to which interpretation is correct. The consequent of these differences is a focus on allowing knowledge to be collaboratively corrected, but little work on how to assist in that resolution. This accurately models the Web 2.0 paradigm and was shown to work in the ICT support domain, but is impractical for a domain such as medicine where conflicts in knowledge may appear without obvious solutions, especially without wide ranging collaboration.

2.3 Data Analysis

The central component of any knowledge discovery method will be the data analysis: the methods which are used to identify relationships, trends, or any other information from the raw data available. This component is covered by steps three and four of Kurgan and Musilek's generic knowledge discovery model described in section 2.1.2 (Kurgan & Musilek, 2006). Clearly, these methods form an integral part of knowledge discovery, as they are the methods which are used to discover what can be gleaned from the data. However, it is important to remember that the result is typically just more data, or metadata; it comes with no meaning or explanation attached; and has no guarantee that the discovered patterns are at all applicable outside of the set of data that was examined. The data analysis therefore

can only be considered to be one step in a knowledge discovery process: the goal of knowledge discovery is to find *knowledge* rather than just extra data, and it is necessary to interpret the new data to determine what knowledge may be concluded. Therefore, while the data analysis process can be considered to come under the coverage of knowledge discovery, it can only constitute a component of that process (Brachman & Anand, 1996; Fayyad, et al., 1996b; Goebel & Gruenwald, 1999).

There are two terms that are used, at times interchangeably, to describe those data analysis methods: *data mining* and *machine learning*. The subtle distinctions between them lie mostly in the desired goal of the analysis, and partly in the manner in which the method performs: data mining methods seek to analyse data to discover trends, patterns, or relationships within the data; machine learning methods seek to analyse data to determine a model for the interpretation of the same or similar data. The difference is often only a semantic one: a pattern or relationship can be used as a model and a model can be considered as a description of trends; hence the common interchangeability of terms. However, the distinction is still made; and the names themselves infer another perspective on the difference: a data mining method is one which trawls through data, attempting to unearth some new data; whereas a machine learning method considers the data, and learns patterns with which to interpret this and other such data.

2.3.1 Data Mining

Data mining has been described as any process for analysing sets of data and attempting to extract some information from them (Witten & Frank, 2005), or as any method of extracting patterns from data (Goebel & Gruenwald, 1999). Typically these methods attempt to find statistical trends and patterns within large amounts of data, which can be assumed to indicate some underlying connection or event. The desired outcome is to be able to describe, with confidence, trends that can be used to predict future actions or events, and use this to some advantage (Fayyad, et al., 1996a).

As a field, and as a term, data mining was initially driven by statisticians and data analysts (Fayyad, et al., 1996b). Over time, ~~"p~~ure" data mining approaches and methods have become less widely used due to the size and complexity of the data

now being mined (Fayyad, et al., 1996a), and the demand for more meaningful output (Goebel & Gruenwald, 1999). To address these problems data mining has become increasingly more integrated with knowledge discovery and machine learning.

2.3.2 Machine Learning

When considering acquiring knowledge, there are methods which can constitute an overlap between knowledge acquisition and data mining: while knowledge acquisition methods focus on extracting knowledge from a human expert, there are also methods which attempt to acquire knowledge through automated means (Witten & Frank, 2005). Machine learning as a term describes any method whose goal is for a computer system to obtain new knowledge about a subject, in a reproducible format, from a set of data. While this incorporates the goals of knowledge acquisition and has an obvious overlap with knowledge discovery, machine learning methods involve the computer taking a much more active role in the learning process: the focus is on how the computer system can identify relevant information about the subject, with minimal human input. These methods take the approach that the reduction of human input is the key to avoiding the knowledge acquisition bottleneck, and also facilitates the removal of unintentional bias, to allow purely statistical and logical methods to find points of interest in the data (Grefenstette, Ramsey, & Schultz, 1990; Hong, Wang, Wang, & Chien, 2000).

2.3.2.1 History

Machine learning as a field came into existence largely because of perceived shortcomings with knowledge acquisition (Grefenstette, et al., 1990). While knowledge acquisition methods showed success in some applications, research and development in the expert systems area discovered that the most significant problem faced, negatively impacting on both the effectiveness and cost of creating an expert system, was the knowledge acquisition phase. As has been mentioned, this ―knowledge acquisition bottleneck" caused a change of attitude in the area, shifting the focus from trying to model human expertise directly, towards automated processes of deriving expertise (B. G. Buchanan & Shortliffe, 1984; Grefenstette, et al., 1990; Hong, et al., 2000; Sester, 2000). Machine learning is the result of that

shift. It has been described as the use of statistical analysis of data to derive knowledge about how a domain functions (Witten & Frank, 2005).

The major benefit of this is being able to create an expert system or to derive domain knowledge by analysing collected data, with limited expertise required: removing the necessity of having a human expert in the domain expend considerable time and effort developing and engineering knowledge in the system (Quinlan, 1986; Witten & Frank, 2005). This is of particular benefit in subject domains where an expert's time is quite valuable. Machine learning methods also allow the possibility of discovering the knowledge in a different manner to the way in which the expert would describe it – this may be an advantage or a disadvantage, depending on the domain and the ability of the experts to communicate domain knowledge. For example, the method may discover relationships that would otherwise go unexplored because the current expertise in the field does not suggest any such relationship could exist; or it may be a disadvantage, because relationships may be discovered which are present in the dataset but which are not present in the wider domain. It may also be disadvantageous because the method of discovering the relationships can be less efficient, effective or comprehensible than those used by an expert (Piatetsky-Shapiro, 1990).

2.3.2.2 Machine Learning Drawbacks

Machine learning methods are generally most effective in applications where the data that is being used for acquiring or discovering knowledge is sufficiently detailed that conclusions can be drawn from it alone, without further domain knowledge being applied – typically data that has been classified as being of a certain type, or that can easily be categorised according to type, allows statistical methods to find new relationships from the existing relationships and other data (Witten & Frank, 2005). The existing classifications represent a level of domain expertise that has been applied to the data, either from an expert who has examined each case and provided the classifications as extra information, or from an expert who knows which attributes of the set are important.

Machine learning methods are also only particularly effective in domains where the target knowledge (i.e. the knowledge the method is trying to discover) is relatively simplistic: complex relationships which have a practical use are difficult to derive

without also deriving large amounts of other relationships which are meaningless, coincidental, or overly specific to the dataset (Witten & Frank, 2005). When the goal of the machine learning is knowledge discovery, not just data mining for the purposes of training an expert system, it is required for an expert in the domain to examine the relationships discovered and to determine what is useful and what is not (Abe & Yamaguchi, 2005). If the relationships are too many or too complex then this will be a highly difficult and time consuming process, negating the advantages of this approach.

Another drawback is that machine learning can only discover knowledge that is present within the dataset being used: if the dataset is of insufficient size, or happens to contain statistical relationships which are not representative of the domain, then the method will either miss relationships or find misleading relationships; whereas an expert can use their extended knowledge of the domain to make judgements on what is likely to be coincidence and what is likely to be supported by further data (Hall & Smith, 1998).

2.3.3 Data Mining and Machine Learning Methods

There have been a large number of approaches developed for performing data mining or machine learning tasks. This section will describe some of the key developments in the field.

2.3.3.1 Statistical Methods

The first data mining tools were essentially computerised versions of the existing methods, leveraging mechanical computational power to perform the typically time-consuming, tedious, and error-prone tasks required in data analysis (Tukey, 1977). These initial methods are statistical, mathematical approaches to data analysis; they are still commonly in use for manual data analysis and they form the foundation from which most data mining methods are derived.

There is no need to explain the detail of the more common statistical methods here, which most readers will be familiar with: regression, Student's t-test, the analysis of variance and their many derivations and related methods provide a strong body of data analysis tools which are widely used.

Initially, computational power was applied primarily to improve the speed of these methods. Data mining had its beginnings in simple statistical methods and the fast and accurate visualisation of data, to allow human experts to identify relevant information. This work was described as exploratory data analysis, as the work was often tentative: it was uncertain what would be discovered from the data, so the data was considered in as many different ways as possible (Tukey, 1977).

2.3.3.2 Information Theory

In the late 1940s the rapid development of communication and encoding technologies led to a discussion of a mathematical theory for communication (Shannon, 1948). Through these discussions Shannon's work defined the foundation for the field of *information theory*, a mathematical approach to describing communication from which many advances in a range of fields have been made.

Information entropy is one of the fundamental components of the field described in Shannon's paper (Shannon, 1948). It is a measure of the unpredictability of the content of a message; the central tenet is that the more predictable a message is, the smaller it can be encoded. The formula is shown below:

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

The formula finds the entropy (H) of the set of probabilities $p_1 ... p_n$, where any given p_i is the probability of one possible value (i) for one element of the message being encoded.

For example: if encoding a series of dice rolls, there are six equally likely results for each roll, and therefore the encoding for each roll needs to allow evenly for six possibilities. When all possibilities are equally likely, as in this example, the calculation derives maximum entropy (a result of 1). If the die were weighted such that when rolled it always came up 1, this would have zero entropy: the result needs no encoding at all, as it is completely predictable, and the length of the series (how many times it was rolled) is the only information required. If the weighting were less significant, such that the die came up 1 50% of the time, the entropy value would be reduced as less information needs to be transmitted: fewer bits would be

used to encode a 1 result than the others, as it is known to occur more frequently. As these measurements are based on the probabilities of each possible result, the entropy calculation can be described as the content of the message which can be predicted, based on the biased probability (Witten & Frank, 2005).

This theorem can be extended to include *conditional entropy*, or the entropy of one variable when the value of another is known, or $H(X|Y)$. The conditional probability is calculated by considering the individual entropies of an element X for each value of element Y, weighted by the probability of finding that value for Y. The resultant value measures the average uncertainty of X when Y is known (Shannon, 1948). This calculation can also be used to determine the relationship that exists between two variables: if there is zero entropy for X given Y, then Y can be used as a predictor for X, as knowing the value of Y implies a certain value for X; if there is maximum entropy for X given Y then it cannot be used as a predictor, as the values of X are all equally probable as regards to any single value of Y (Khinchin, 1957).

Information Gain

Information gain is a widely used measure in information theory and in data mining and machine learning (Freitag, 2000; Kent, 1983; Quinlan, 1986), occasionally referred to as Kullback-Leibler information gain after the creators (Kent, 1983; Kullback & Leibler, 1951). It calculates how much the entropy is reduced for a variable X if the value of Y is known. Thus it can be defined as $igain(X|Y) = H(X) - H(X|Y)$, where $H(X)$ is the entropy of the variable X and $H(X|Y)$ is the conditional entropy of X with known values for Y (Kullback & Leibler, 1951). As described previously, a reduction in entropy is a reduction in uncertainty, and a decrease in the cost of encoding; information gain quantifies this reduction, allowing comparisons of the benefits of knowing the value of a variable, or, of how well one variable can predict the value of another.

This trait has led information gain to be applied in numerous data mining and machine learning studies, including Quinlan's popular decision tree induction methods (Kodaz, Özsen, Arslan, & Günes, 2009; MacKay, 1992; Quinlan, 1986). In Quinlan's ID3 approach, information gain was used in building the decision tree to identify the best attribute to "split" on at the current level of the tree, or which attribute provided the most accurate segmentation of the data based on the class

attribute. At any given step, the attribute that is best at predicting the class will be chosen for the current level of the decision tree, by calculating the information gain for the class attribute given knowledge of the current attribute: the largest information gain is the attribute which provides the most extra information about the class.

2.3.3.3 Rule Induction

As researchers and developers struggled with the problems of the knowledge acquisition bottleneck, a diversity of alternative methods were produced in the late 1970s and early 1980s. One of the first and most intuitive methods developed was rule induction. The premise of this approach is that a system can examine data and the outcomes (or classifications) of that data, in order to differentiate what might be the causes for each result (or class). It can then develop rules which accurately classify each case that it knows about.

Waterman established one of the first studies into the area in 1970 (Waterman, 1970), and as the knowledge acquisition bottleneck problem became more prominent the approach was slowly adopted by more of the research community. By 1980 many studies had been undertaken, including some major projects such as DENDRAL (B. Buchanan, Mitchell, & SCIENCE., 1977; Michalski, 1978; Simon & Lea, 1974).

Many methods were developed for inducting rules, but all follow the same general pattern. Once a body of cases have been identified as belonging to the same class, the system can look for the attribute, or attribute-value pairing, which best differentiates those cases from the cases without that classification. This attribute or attribute-value pair becomes the first condition of a rule defining that classification; and the process iteratively continues.

One of the biggest drawbacks of such a system is that in order to know how to create the rules the data needs to have been pre-classified: either having had the classification already added, or for one of the attributes of the case to have been selected as a differentiator, or class-attribute. This then allows the rules to be generated by looking at which other attributes, and which values for those attributes, would seem to indicate certain classifications (Roberto J. Bayardo & Agrawal, 1999). Depending on the number of cases, the number and nature of the attributes

for each case and the complexities of the classifications, this process may be relatively trivial or impossibly time consuming. For example, consider a database of thousands of patient test results, collected over a number of years, which have not had the final diagnoses attached (a surprisingly common occurrence (Roddick, et al., 2003)): a medical expert would have to be employed to examine each and every case, interpreting the results and adding their diagnosis; a process which could take months of full-time work. Note also that the expert is restricted to the data that is available: they cannot request further tests to assist in their interpretation and classification. This highlights a further flaw in the approach: even if the data had been classified as it was entered (i.e. a final diagnosis is recorded with each case), the process could be hindered and mislead by errors or missing values in the data; especially if some data that was used to make the classification is not in fact present in the dataset.

The resultant learned classification rules present their own difficulties: there may be far too many rules for a human expert to be able to examine and determine which rules are valid, which do not describe interesting information, and which are worth considering for further study (Bachant & McDermott, 1984; Barker, O'Connor, Bachant, & Soloway, 1989). The method is also prone to generating very simple rules which provide no real benefit, and very complex rules based on dataset-specific, coincidental relationships (Towell & Shavlik, 1994).

2.3.3.4 Decision Trees

A related field, following a similar pattern of development, is the induction of decision trees. Decision trees make a logical choice for simple automated learning, as, generally speaking, they are developed incrementally: at each point, the system only needs to identify one differentiating factor, which then splits the data into smaller segments; this process continues until the data is separated into distinct classification groups. As induction of rules provided successful results and the knowledge acquisition bottleneck became an increasing concern, induction of decision trees gained impetus. As mentioned previously Quinlan's C4.5 algorithm for the induction of decision trees, and the subsequent improvements and extensions to the method, were very successful and popular in the late 1980s (Quinlan, 1993).

Induction of decision trees began with Hunt's concept learning system framework (CLS) in 1966 (Hunt, Marin, & Stone, 1966). This method took the approach of selecting decisions that resulted in the minimum cost of classification: including the costs of identifying the value of the attribute and the cost of misclassifying a case; and made these calculations to a variable number of steps ahead, in order to construct better overall tree paths (Quinlan, 1986). This system proved effective, but potentially computationally slow depending on how far ahead it calculated. This led Quinlan to develop the ID3 algorithm, which removed the look-ahead cost calculation in favour of an information theory calculation, focussing on identifying the optimal decision for the current step (Quinlan, 1979, 1986). Until this point, all algorithms had required the use of discrete-valued attributes, but Paterson and Niblett adapted the ID3 approach to allow the use of integer attributes (Patterson & Niblett, 1982; Quinlan, 1986); this was later generalised to any continuous-valued attribute in the ASSISTANT study (Kononenko, Bratko, & Roskar, 1984; Quinlan, 1986). Quinlan's decision tree methods made use of the information gain calculation, to determine the split condition that will accurately assign the classifications for the most number of cases, in the current set. The gain is based on comparing the accuracy of the system after the split, for each of the subsets of cases, compared to the accuracy before the split (when all cases were assigned to one category) (Quinlan, 1986).

After C4.5 was developed many further extensions were developed and applied, as there were a number of disadvantages to the approach. As with any classification method learned from the data, decision trees often have difficulties with generalisation: they may learn to classify the cases which they have seen perfectly, but the conditions chosen may represent coincidental relationships rather than real relationships; even if the conditions do represent legitimate domain reasoning, the values used can only be based on the examples that are available, and so are unlikely to accurately define the classification (Quinlan, 1987; Witten & Frank, 2005). A similar problem is that trees can easily become *over-fitted* to the data: that is, they overly focus on the details within the data, and so cannot be generalised to other cases (Davison & Hirsh, 1998). Some very successful extensions were developed to overcome this issue, under the name of *pruning*. The general idea is that once the tree has been developed some of the lower branches should be

removed, as these will likely be providing specific distinctions between individual cases in the dataset and will not be representative of general domain trends (Quinlan, 1993). A successful and widely adopted approach is *reduced error pruning*, whereby the dataset is divided into a training set and a test set. The training set is used to build the decision tree as normal; the test set is then used to test the accuracy of the tree, as it is iteratively pruned. Each non-leaf sub-tree of the current tree is tentatively replaced by whichever leaf is most correct. If the modified tree gives equal or fewer errors, then the modification is kept (Quinlan, 1987).

2.3.3.5 Case Based Reasoning

Another group of methods which finds application in both knowledge acquisition and in data analysis are case based reasoning methods. These methods can be adapted to a machine learning approach by using pre-classified data, or by using one of the attributes of the case as the class variable (Watson & Marir, 1994).

An example of one of these approaches is the *k-nearest neighbour* (KNN) method, which was first established in 1951 (Fix & Hodges, 1951), and further formalised 10 years later (Johns, 1961; Witten & Frank, 2005). KNN is a method founded on case based reasoning principles for classifying cases based on their similarity to other cases. This similarity is defined as a hyper-dimensional distance metric: that is, if the cases were plotted as points on a hyper-plane, with one dimension for each attribute, the distance between two cases is used to identify their similarity. A new case is given the same classification as the majority of the *k* cases which it is closest to on the hyper-plane; in other words, its *k* nearest neighbours.

This algorithm learns incrementally, potentially increasing its knowledge with each new case examined without requiring all data to be re-evaluated. However as the number of cases seen increases, or the number of attributes in each case increases, so too does the efficiency of the method substantially decrease (Kurniawati, Jin, & Shepherd, 1998). The method also has difficulty defining complex relationships, except through the storage of a sufficiently large number of cases and the examination of a sufficiently large number of attributes, such that the complexity can be accounted for and any relationships found. However, as noted, this can cause the system to become highly inefficient. A further problem is that the knowledge gained from storing these classification boundaries is not easily viewable and

understandable by a human expert, due to the potentially vast number and multidimensional nature of the spaces being defined (Hand & Vinciotti, 2003; Kurniawati, et al., 1998).

2.3.3.6 Clustering

Clustering describes a methodology of attempting to identify patterns and trends in data by finding what case groupings exist, and how those groups are defined (Jain, Murty, & Flynn, 1999). By performing this task, classifications can be derived that were previously unknown, and a simple comparative analysis of the members of each cluster will provide conditions for determining cluster membership for future cases. This ability to find truly new classifications makes clustering a powerful data mining tool, although obviously it should be noted that any new classifications discovered have no associated meaning: they simply provide evidence that classes exist within the data, and it is an expert's responsibility to determine why this should be true and what the implications of this are (Jain, et al., 1999).

k-Means Clustering

k-Means clustering is one of the most simple clustering methods, and many clustering methods use the k-means approach as a template (Berkhin, 2006). k-Means finds results by many repeated passes of the same function: each case in the dataset is assigned to the cluster that it is closest to, based on a hyper-dimensional plot of all cases, with each attribute in the dataset describing one dimension. "Closeness" is a complex term in clustering, and is where most of the differentiation between methods lies: in k-means, closeness is decided by comparing the average, total, or maximum (depending on implementation) difference between the mean of all cases currently in the cluster and the current case under consideration (Hartigan, 1975; MacQueen, 1967). Initially, the clusters are decided by randomly assigning one case from the dataset to each cluster. Once all cases have been assigned to a cluster, the process is repeated, with the mean-points of each cluster constituting the midpoints of the new empty clusters. This is repeated until there is no deviation of cases from one cluster to another between successive runs; at which point, the clusters are determined to have stabilised and the results presented (Hartigan, 1975; MacQueen, 1967). The number of clusters (k) that are initially created is determined by the person running the clustering method. If this is not known, the clustering

process can be run multiple times with different numbers of clusters, to attempt to find the best results (Hartigan, 1975; MacQueen, 1967).

Clustering Limitations

The biggest flaw with k-means, and with most clustering methods, is that while they can find completely new class groupings the method requires that the user input how many of these groups to look for (Hartigan, 1975; Witten & Frank, 2005). This requires some level of understanding of what the results are likely to be before the process is run – severely dampening the benefits of discovering new class groupings. This is exacerbated by the second major flaw with clustering methods: that they are very expensive in terms of time and processing power, particularly if the number and nature of the clusters being looked for is uncertain (Hartigan, 1975; Witten & Frank, 2005). Clustering works quite effectively and relatively efficiently with a small set of attributes, cases and clusters as parameters; however as these numbers rise the processing time dramatically increases, in many cases to the point of being unusable (Witten & Frank, 2005). This time requirement can be reduced the more that is known about the clusters being searched for: restricting the search space to a small number of attributes, or weighting important attributes more than others, will dramatically improve the speed and efficiency of the process; as will reducing the cases being examined, or specifying the approximate number of clusters to search for.

Clustering methods are usually non-deterministic, with the assignation of random cases to the initial clusters determining how the final clusters will be formed: however, by having stringent stabilisation requirements it is generally assured that if there exist clusters within the data being analysed, they will be discovered (Jain, et al., 1999). This is still a downfall of the method however, as there is no guarantee that the results are the best possible results, and there will always be doubt.

A further problem is that clustering can often be inconclusive, as methods generally provide little distinction between obvious, strong clusters and weak clusters. The clustering algorithm only functions to the extent of having stabilised and defined clusters: the veracity of these clusters over larger amounts of data and how reliably they can be defined is not presented: this must be pre-determined by considering the domain, and concluding whether strong clusters are likely to exist or not (Jain, et al.,

1999). The end result of these drawbacks is that for a clustering method to be effective there usually needs to be a significant level of expert involvement and application of domain knowledge: without this, a clustering method becomes a blind search, likely to take significant time only to discover clusters which are uncertain and uninformative. This unfortunately means that while clustering can find new knowledge and entirely new classifications, to do so effectively requires that much of the nature of the classifications is already known. Clustering is therefore a method which lends itself to quantifying known relationships, or relationships that are expected to exist, rather than a method of discovering new knowledge.

2.3.3.7 Association Rule Mining

In the early 1990s, as the technology for recording business data became increasingly prevalent, there began an increase in studies for analysing that data: beginning the field of data mining. One of the earliest applications of this was to identify shopping trends in large-scale databases of customer transactions. Each record in these databases constitutes the list of items that the customer bought in that shopping transaction. This problem presented unique challenges in that each case did not conform to having the same limited set of attributes and a value for each one, but rather had a variable-numbered combination of a large possible item set. Nevertheless, in order to work with these cases each was typically treated as having a large number of Boolean attributes, one for each possible item: most of which will be false for each case. This allows the automatic identification of which items are associated with which other items; in this case for identifying how better to market products, but can be generalised to discovering which attributes are related to which other attributes (Agrawal, Imielinski, & Swami, 1993). The resultant association rules are of the form $X \rightarrow I$, where X is an item (or attribute) set, and I is an individual item (or attribute) which is not contained in X (Agrawal, et al., 1993).

Interestingness Measures

While it is a trivially simple task to define all the possible association rules that could be supported by a dataset, or more correctly a data model, the difficulty lies in identifying which of those rules are adequately evidenced by the data to warrant

regarding them as reasonable and indicative of knowledge (Agrawal, et al., 1993). The simplest of these, *support* and *confidence*, form the fundamental components of many of the other measures, some of the more common of which are described here.

The *confidence* of a rule is the percentage of the cases that have the antecedent of the rule (*X*), that also have the consequent of the rule (*I*); for example, in a dataset of 20 cases, if 10 cases had all items in *X*, and of those 10, 4 cases had the consequent *I*, the rule confidence would be 40%. The confidence suggests how likely this rule is to represent a true association: if 100% of cases with item *A* also have item *B*, the system can be maximally confident that there is an association between *A* and *B*; whereas if only 5% of cases with *A* also have *B*, this is much more likely to be coincidental, or at least unreliable enough not to warrant further action.

Support attempts to describe the statistical significance of a rule: it is the fraction of cases in the dataset that satisfy the rule, having both the antecedent and the consequent. This helps to indicate the likelihood that an association is able to be generalised beyond the current data (Agrawal, et al., 1993).

Lift, also called *interest* (Brin, Motwani, Ullman, & Tsur, 1997; Roberto J. Bayardo & Agrawal, 1999), is a measure of how singularly dependent the consequent is on the antecedent. A low value for lift indicates that the consequent is unlikely to be dependent on the antecedent. *Lift* can be defined as (Roberto J. Bayardo & Agrawal, 1999), (Brin, et al., 1997):

$$lift(X \rightarrow I) = \frac{confidence(X \rightarrow I)}{support(X \rightarrow I)} = \frac{\mathrm{P}(X, I)}{\mathrm{P}(X)\mathrm{P}(I)}$$

In more literal terms, the *lift* value describes how much more likely, multiplicatively, the consequent is to appear in the set of cases that have the antecedent, than in the overall set of cases. For example: as previously, there are 20 cases, 10 having *X* and 4 of those with *I*, giving $X \rightarrow I$ a confidence of 40%; but of those 20 cases the only that have *I* are the 4 that also have *X*; then the rule has a *lift* of 2. Although only 40% of cases with the antecedent *X* have the consequent, which would seem to indicate a relatively weak correlation, *X* is still twice as good a predictor of *I* than random selection, which can only predict cases which have *I* 20% of the time. *Lift* is also a symmetrical measurement, in that:

$$lift(X \rightarrow I) = lift(I \rightarrow X)$$

Gain is a measure used by Fukuda et al to help find optimal ranges for rule definition, and is defined as:

$$gain_\theta(X \rightarrow I) = support(X \rightarrow I) - \theta \times support(X)$$

where the variable θ represents the minimum confidence threshold (Fukuda, Morimoto, Morishita, & Tokuyama, 1996). Explicitly, the resultant value of gain is the number of cases that support the rule above the minimum necessary for the rule to match the confidence threshold, given the support for the antecedent. Continuing from the previous examples, if the minimum confidence threshold was set at 20%, then $gain(X \rightarrow I)$ would be 2: as the minimum number of cases required for the confidence to meet the threshold of 20% is 2 ($0.2 \times support(X)$), and the $support(X \rightarrow I)$ is 4.

Piatetsky-Shapiro defined a further interestingness measure in 1991, which Bayardo and Agrawal pointed out is a specialised case of *gain*, with θ fixed as $\frac{support(I)}{|D|}$, where $|D|$ is the number of cases in the dataset (Bayardo Jr. & Agrawal, 1999; Piatetsky-Shapiro, 1991). Thus the measure calculates: of the cases that support the antecedent, how many more have the consequent than would be expected, using the ratio of number of cases with the consequent against the dataset to derive the expected value. To illustrate, again using the previous examples: if 10 out of 20 cases have antecedent *X*, and 4 cases have consequent *I*, all 4 of which also have *X*, then the *p-s gain* is 2; which is indicating that the antecedent *X* appears in association with the consequent *I* in 2 more cases than would be expected. 4 cases have both consequent and antecedent, while based on the ratio of cases with the consequent to cases in the dataset (0.2), it would be expected that only 2 of the 10 cases with the antecedent would also have the consequent.

Conviction is another function of *confidence* which was designed to complement *lift*, as it considers the probabilities of both the consequent and antecedent individually (Brin, et al., 1997). It is defined as:

$$conviction(X \rightarrow I) = \frac{|D| - support(I)}{|D|(1 - confidence(X \rightarrow I))}$$

Each of these measures provides a relative measurement of interestingness for each possible rule; however as they are based from different measurements they will

often provide conflicting rankings. In order to improve the efficiency of the data mining, all measurements for all possibilities are rarely calculated: rather, a first pass is run finding all rules which match a minimum threshold for simple measurements such as confidence and support, then more complex measurements made over the remaining rules (Agrawal, et al., 1993; Bayardo Jr. & Agrawal, 1999; Lenca, Vaillant, & Lallich, 2006). These will again often have minimum thresholds, displaying only those rules which surpass the threshold value for each measure. The literature generally does not suggest optimal thresholds for these or other measures, and threshold values are rarely discussed in detail. The most common view is that the thresholds should be modifiable by the user, as the required minimum interestingness of a rule is dependent on the data and what the user is looking for (Hidber, 1999; Lenca, et al., 2006; Tan & Kumar, 2001), although some methods have attempted to develop relative thresholds (Lavra , Flach, & Zupan, 1999).

The major problems with association rules are the computational complexity of identifying the rules and the often vague results: the method provides absolutely no explanation for why these associations exist, which makes it difficult to quantify exactly how well an association might generalise, or what to do with the associations once discovered. Nevertheless association rule mining became a very popular approach in data mining which found wide application in marketing research (Fayyad, et al., 1996a).

2.3.3.8 Neural Networks

Neural networks are an approach to developing a self-learning system, based on our understanding of the fundamentals of the human brain. As with the brain, a network of neurons is established through which information passes, causing some neurons to fire; and the extent to which each of the neurons fire determines what the final output of the network is. The *perceptron*, the first method which could be called a neural network, was developed in the 1960s based on linear regression techniques. This method calculated appropriate weights for each attribute such that the sum of the weighted values could be used to predict the class of each case (Nilsson, 1965; Witten & Frank, 2005). It was soon identified, however, that this method had fundamental limitations (M. Minsky & Papert, 1988), and research in the area did

not regain popularity until more complex designs were developed (Witten & Frank, 2005), such as the multi-layer perceptron and *backpropagation*.

In general, the construction of a neural network an initial set of neurons are established, each of which take different elements of the case as input and have different functions deciding whether they fire, given a range of input values. These initial neurons can then feed into another perceptron layer, which potentially feed into further layers, with neuron outputs becoming the inputs for the subsequent layer. Once established, the network learns through an extensive training process of data examination, updating the neuron functions to be more correct for each data instance (Gardner & Dorling, 1998; Witten & Frank, 2005). Many neural network approaches also employ the *backpropagation* technique whereby cases are presented to the network, and the outcome observed; if the outcome is correct the neurons that contributed are positively reinforced, and if the outcome is incorrect they are negatively reinforced (Gardner & Dorling, 1998). In some approaches, layers of neurons can also be added and removed during the learning process (Fritzke, 1993; Huang, Saratchandran, & Sundararajan, 2005). This process continues until the network has stabilised and is returning accurate results.

Neural networks are interesting because they can require very little input or supervision, although they require an initial specification of what a correct result is in order to begin learning. Successes have been made, particularly for problems with *noisy data* (data that contains many errors) that make human expertise difficult to apply (T. Mitchell, 1997). However neural networks are not suited to all domains: they are slow to train, do not learn well from complex data sets, and cannot learn incrementally. Extensive training is required to produce an accurate system, and that system is not adaptable to new situations. Perhaps the biggest drawback is that, even more so than with other machine learning techniques, any knowledge in the system is entirely opaque: it is stored implicitly within the configuration of the network, and the system can give absolutely no justification for why it reaches the results it does (Pernin, 2008; Towell & Shavlik, 1993).

2.3.3.9 Bayesian Classifiers

Bayesian belief methods, or Bayesian networks, are a probabilistic method which produces models of the probabilistic relationships between attributes (Goebel &

Gruenwald, 1999). The core of all Bayesian approaches is Bayes' theorem, which defines the probability of a hypothesis (in this case A) given evidence (B):

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Given a set of pre-classified data, this theorem allows the calculation of the probability of a classification, given certain attributes of a case. This can easily define a classifier: the pre-classified dataset provides the probability of each outcome for each attribute value; and so any new case can take the probabilities associated with the values it has, and find the probability it has for each possible classification. This is called the *Naïve Bayes* approach, as it assumes that each attribute is independent: interdependency between attributes changes the resultant probabilities.

There are other applications of Bayes' theorem: many more complex and specialised methods have been developed based on the principle (Szarfman, Machado, & ONeill, 2002). A more common example is a Bayesian network: this consists of a directed, acyclic graph where nodes are attributes and edges are dependencies between attributes, with probability functions associated with each node describing the likelihood of each outcome for the attribute given the values of the attributes it is dependent on (Heckerman, 1995). This provides a visual representation of the data, and the possibility of structuring attribute dependency into the calculations.

Bayesian networks provide an option for including, and to an extent rely upon, the input of existing knowledge in the form of the dependencies between attributes. There are methods to automatically derive this information, by generating a large number of different topologies and identifying which is the most appropriate. The graph is then updated by analysis of the data, to find more accurate probabilities and possibly new dependencies (Heckerman, 1995).

2.3.3.10 Genetic Algorithms

Genetic algorithms are a similar field of development to neural networks. Where neural networks are modelled after our understanding of the human brain, genetic algorithms are modelled after another proven natural system: genetic evolution. For

this reason it is sometimes also referred to as *evolutionary computation* (M. Mitchell, 1998). Genetic algorithms function by randomly generating a number of possible solutions to the problem under consideration. A subset of these, comprising the most successful strategies with the possibility of some random selections, is chosen for reproduction. The components of the subset are duplicated and combined in some way into a new population, again with some potential for random development; and the process is repeated with that population, until a suitable solution to the problem is discovered (M. Mitchell, 1998).

Evolution has been used for a long time in computing research, with examples as far back as the 1950s (Barricelli, 1957; Friedman, 1959; M. Mitchell, 1998). Much of this initial work was begun out of a desire to test evolutionary models, where many generations of work could be calculated in a realistic time frame, rather than in solving problems through an evolutionary approach (M. Mitchell, 1998). Genetic algorithms were formalised in the 1970s by Holland, as a way of describing the evolutionary processes that he had been modelling (Holland, 1975). However, despite having been in relatively consistent use since their development, genetic algorithms were considered an esoteric and impractical solution, largely due to the high computational cost of finding effective results and the opacity of the processes involved and the solutions found. With the increase in computing power and its availability of the 1980s and 1990s, evolutionary algorithms became more popular, particularly with some commercial successes and a general appeal that many other methods lack (Bentley & Corne, 2002; Thearling, 1998).

Genetic algorithms have been shown to be quite effective at developing classification models, comparable to any other approach (Orriols-Puig, Casillas, & Bernadó-Mansilla, 2008). However, they also have similar drawbacks to neural networks: they are very slow to develop, requiring considerable computation time; they generally require extensive initial configuration to achieve effective final results, but are relatively inflexible during development; and the end result can be quite opaque, unable to provide an explanation for how it achieves its results (Orriols-Puig, et al., 2008).

2.4 Results Analysis

An important component of any knowledge discovery method is to analyse the results that have been obtained (Fayyad, et al., 1996b). This step is often not given significant consideration when developing a knowledge discovery method: as knowledge acquisition and data mining are fields unto themselves, they retain most of the focus, with little work on combining the methods (Fayyad, et al., 1996b; Piatetsky-Shapiro, 1990; Pohle, 2003; Sinha & Zhao, 2008). There are thought to be a few reasons for this lack of focus. Results analysis is often considered to be outside the scope of those developing and deploying the knowledge discovery method, as it is thought to be more productive and quantifiable to focus on modifying the data mining method to produce less results, and try to ensure that it produces only good results (Freitas, 1999; Matheus, et al., 2002). Results analysis is also one of the most difficult to specifically develop for (McGarry, 2005). The standard method is to simply present the results to an expert, and have the expert identify which are useful and which are not (Matheus, et al., 2002; Pohle, 2003).

It generally requires someone with significant expertise and experience to recognise when newly discovered knowledge is reliable, or is indeed useful (Brachman & Anand, 1996; Fayyad, et al., 1996a). While an expert system may well be developed which can reproduce the knowledge of an expert for any case it could be presented, this expert system will only be an expert on the things it has been taught: it may have much of the knowledge that the human expert has, but it lacks the ability to adapt and apply that knowledge to new situations. This key fact has hindered the development of effective automated analysis devices in knowledge discovery research. It is also useful to consider that it is very likely that only an expert in the field will be able to put new knowledge to any use: and so it is necessary to involve such an expert in identifying the accuracy of the results (Fayyad, et al., 1996b).

The ability to interpret results and find knowledge is dependent on the method used to discover them: the patterns found from a neural network are very difficult to interpret, whereas the patterns learned from classification rules are much more easily examined and checked by a human. When trying to discover useful new knowledge, the choice of method used, and the type of results it produces, is therefore an important consideration (Piatetsky-Shapiro, et al., 1994).

2.5 The Medical Domain

The word *domain* is commonly used in expert system development to mean the field to which the system is being applied, or in other words, the things which the knowledge and data describe, and which the system is intended to gain expertise about. The methods developed in this study have been tested and are intended to be applied in the medical domain, and while perhaps they need not exclusively be used for this purpose, they have been developed so as to work best with this sort of data. For the purposes of this study, the medical domain is defined as those fields of work and research that involve the functions of the human body. These fields are characterised by a number of common traits that complicate work in the area and require specific consideration.

2.5.1 Medical Knowledge Discovery

The medical domain is considered one of the most difficult domains for knowledge discovery, but also one that can provide significant benefits (Roddick, et al., 2003). Unlike some other domains, the benefits of discovering new medical knowledge are obvious: and there is arguably no greater goal than the improvement of health and medical understanding.

2.5.1.1 Difficulties in the Domain

Although the discovery of new medical knowledge is a desirable goal, there are a number of factors which make it a very difficult task. The domain presents a distinctive combination of challenges for study and practice, which are outlined here.

Medical Data

One of the most significant difficulties faced in medical fields is the problem of gathering accurate data. Most of the functions being measured are internal to a human body and may be invisible to the human eye; and by definition, all functions involve elements integral to the patient's health. This typically makes direct measurement an impossibility, such that most data consists of measurements incidental to the events being considered (Cios & Moore, 2002a). Often the tests being performed to generate the measurements are still intrusive in some fashion,

and extrapolation to a normal state is still necessary. These factors mean that medical decisions are often made based on best interpretations and estimates without the solid data that other fields can provide (Cios & Moore, 2002a).

Given the difficulty in measuring the desired functions and components, there is a large field of research and development devoted to finding new tests that can provide different and better information about the functioning of the human body. Combined with the endless development of making tests cheaper, faster, and easier, these factors result in each case having a plethora of different measurements and test results to analyse and make decisions from. When this is multiplied by the ever increasing number of fields of study within medicine, the amount of data contained in an individual case can become overwhelming (Pribor, 1989).

Frequently the data is also incomplete: in many situations it is unlikely that a final diagnosis be recorded with the case, which may or may not have all test results included in a central location; it seems quite common that cases are recorded with no conclusions or interpretations added, and where they are recorded, they very rarely conform to any sort of standard (Roddick, et al., 2003).

This difficulty extends beyond the immediate construction of the data through measurement, to ethical and social issues regarding how the data is handled. The data is fragmented between different hospitals, clinics, general practitioner surgeries, and government departments; and often only weakly linked via non-standardised identifiers. When this data is accessed by individual projects, for ethical and legal reasons the data is always de-identified, making useful linking to other data a difficult, if not impossible task. This also requires a greater level of security than other domains, and further restricts the ability to compile useful datasets (Cios & Moore, 2002a). Although the data is collected widely and in great numbers, there is often little collaboration between the points of collection; and at times organisations are openly distrustful. Determining ownership of medical data is difficult and controversial: whereas ownership is typically tied to the right to sell an item (Cios & Kacprzyk, 2001; Cios & Moore, 2002a), legal and social restrictions on the selling of medical data confuse the issue. Given highly publicised legal action on such matters, and no resolution to determine ownership in general, institutions are often hesitant to share data at all (Cios & Moore, 2002a).

Medical Knowledge

The medical domain is an expansive field of study and practice, and contains a vast array of knowledge. This knowledge covers all the many aspects of the human body and how it functions, and the many, many other organisms, effects and problems that can interfere with that function. This knowledge needs not only to cover how all these factors work and inter-relate, but also how to identify them, how to measure them, and how to treat them.

The indirect nature of the data mentioned in the previous section contributes to the quantity of knowledge in the domain, as all medical experts are required to understand how to interpret the implications of measurements incidental to the event under consideration. Given the uncertain nature of most medical practices, such as interpreting a series of symptoms and test results to form a diagnosis, medical experts need to be able to identify all possible causes of the data that they have; and know what further tests to perform to eliminate each possibility, until a single possibility remains. Thus medical knowledge by necessity relates to all possible causes of all possible data values, and which tests and measurements can be used to discover each of them. It also commonly presents a different paradigm to that in any other domains. The null hypothesis approach is much more frequently applied in medicine than other areas, as knowledge often describes how to eliminate possible conclusions rather than how to identify them (Roddick, et al., 2003).

The complexity of this knowledge impacts on medical knowledge discovery tasks: it necessitates the inclusion of domain experts more than many other domains, as while a knowledge engineer may be able to identify relevant results in a marketing system, or be able to make rudimentary decisions about the data and knowledge to be included, this often cannot be said for medical systems; the additional education and training required to cover the breadth and complexity of medical knowledge is too great. Thus any knowledge discovery project involving the medical domain needs a strong collaboration between the engineers and the medical experts (Roddick, et al., 2003).

Incompleteness of Knowledge

To compound the difficulties with gathering and understanding the data, the knowledge of exactly how the body functions is incomplete, and is constantly being

revised, updated and added to. While understanding has greatly improved and continues to improve with further study, the functions of the human body still provide many mysteries which require yet further study. For these reasons, medical study is unlikely to ever be fully complete, and will always require further work and study (Fox & Bennett, 1998; Steeves, 1965).

Disagreement between experts

A further complication in acquiring and using medical knowledge is that the most recent, and sometimes most relevant, knowledge is often not well defined. This problem leads to, and is caused by, alternative views between experts. A major cause of these disagreements between medical experts is that much of their knowledge is learned through experience rather than from a central source. This is due to both the nature of the training, knowledge, and data in the domain. The data consists of indirect measurements which are not guaranteed to be representative of a single event, problem or state; the data can also vary widely from patient to patient, for a huge number of reasons due to the astonishing complexity and variability of the human body. As a consequent of these difficulties, expert medical knowledge consists of interpretations of these variable, incidental measurements: and in many cases these interpretations will be ambiguous and inconclusive. Contributing to this is that the experts who have the most practical experience will often not have the most training and theoretical study (unavoidably, as there is only so much time in each person's life); and so, a method of interpretation that one expert may have learned from published studies of others, another may have learned through first-hand experience. This leads to inevitable discrepancies as one considers the differences inherent in learning through experience and learning through published works; the massive variability that can be present between populations of patients; and the natural effects of misunderstanding and errors of judgement that can lead conclusions to be made erroneously. As each expert will have formulated their opinions based on their own evidence, there is typically little indication of which, if any, of the options or methods are correct; and correctness usually becomes of little concern, provided the knowledge can be shown to be used effectively and without inappropriately high levels of risk.

2.5.1.2 Computational Studies into Medical Knowledge

Expert Systems

There is a rich history of computation being applied to medical knowledge. One of the earliest expert systems, MYCIN, was a rule-based system developed in the early 1970s for identifying bacteria and recommending antibiotics (Horvitz, 1986). MYCIN was one of the key studies that established the expert systems field, as it was a large and complex system successfully deployed. Since then, many expert systems have been developed for the medical area, and continue to be developed: for a few examples, see (Aikins, Kunz, Shortliffe, & Fallat, 1983; Edwards, Compton, Malor, Srinivasan, & Lazarus, 1993; Pribor, 1989; Shortliffe, 1974; Singh, 2006; Snow, Fallat, Tyler, & Hsu, 1988; Stansfield, 2009). The popularity of expert systems in this field is unsurprising, since the volume of data to be considered, and the complexity of the tasks involved, make computerised systems a logical choice (Pribor, 1989). Additionally, the consequences of error mean that any additional support, which may improve the quality and consistency of service, is beneficial. However, the complexity of the data, the complexity of the knowledge, and the high accuracy required contribute to making medical expert systems development a difficult process: and hence a problem which attracts expert systems researchers. The applications are also quite varied, including diagnosis, prognosis, data interpretation, and education (Masić, Ridanović, & Pandza, 1995).

Data Mining and Knowledge Discovery

As with expert systems, there is a strong history of data mining and knowledge discovery in medicine. The medical domain was identified as an excellent field for data mining early in the history of the field (Piatetsky-Shapiro, 1990), and the two have since been tightly linked (Abe & Yamaguchi, 2005; Agahi, 2007; Cios & Kacprzyk, 2001; Cios & Moore, 2002b; Kononenko, et al., 1984; Prather, et al., 1997; Roddick, et al., 2003; Tsumoto, 2004; Tsumoto & Tanaka, 1996). For the same reasons that the reproduction of domain knowledge is beneficial, the discovery of new domain knowledge is also useful. Unfortunately it also suffers from the same controversies and difficulties.

The issues that apply to expert system development and knowledge acquisition also apply to knowledge discovery, as the complexity and breadth of current medical knowledge mean that any attempt to discover new medical knowledge requires a substantial input of existing knowledge. As an example, one study which did not make allowances for existing knowledge found that very nearly all discovered knowledge was already known to the medical experts involved, prompting future work to involve a redevelopment so as to include an extendable knowledge base (Gialamas, et al., 2003; Roddick, et al., 2003).

However, analysis of data by experts and medical researchers is often still a primitive process. Specific questions may be answered by specific studies, which has a considerable cost in time, money and other resources, and requires the recruitment of subjects to provide data; despite many institutions spending extensive resources on maintaining vast databases of patient information. Unfortunately this data is often only accessed for specific projects, and following the enlistment of a computer scientist who might have their own development goals. For many studies, even with specific software development, data analysis is performed manually with relatively primitive tools (Agahi, 2007).

2.5.2 The Lung Function Domain

The study and medical treatment of the respiratory system is a typically complex medical field. The purpose of the lungs is to facilitate the assimilation and exchange of gases between the atmosphere and the haemoglobin in the blood: specifically, to take in air to oxygenate blood, and to remove carbon dioxide from the blood and expel it (Hughes & Empey, 1981). The lungs themselves are made up of many components, including the airways, alveoli, pulmonary blood vessels, respiratory muscles, and other respiratory controls, all of which contribute to this process (Ruppel, 1994).

Lung function is an important field of study and practice, dealing with one of the most critical components of the human body, without which a human cannot continue to live. The general functioning of the lungs is well understood, but as with most medical fields there is still much that is unknown and requires further study: understanding exactly how different aspects of the lung work under different circumstances and for different people; determining how diseases affect the lungs,

and how best to prevent and treat them; and the problems of new and adaptable diseases provide an endless course of study and development. Being of such vital importance to life, study into the lungs and how best to identify, prevent, and treat problems is of a high priority, with many medical practitioners and researchers employed as specialists in lung function (Cotes & Leathart, 1993).

2.5.2.1 Lung Function Experts

There are many levels and distinct forms of expertise in lung function. While there is certainly overlap between the two, the biggest divide is between respiratory clinicians and technicians. Respiratory clinicians primarily practice lung function testing and interpretation in the context of a clinic or hospital, for the purpose of treating individual patients. Technicians primarily perform lung function testing and interpretation in the context of the laboratory, with an end point of classifying a patient's lung function test results rather than determining their treatment.

In addition to this divide of application of knowledge, there are divides among lung function experts between levels of knowledge. While it is natural in any field that some people will have more training and experience than others, it is in an established aspect of the medical domain that the many occupations which require medical knowledge each require many types of expertise, to differing levels of complexity. For example, a doctor working in an emergency room, or a general practitioner (GP), will need at least a basic level of understanding of almost all medical fields as they may encounter and be required to diagnose and treat any combination of medical problems. Similarly, a nurse will require some understanding of many different medical fields. Depending on the location, the specifics of the person's training and previous employment, the professionals in each of these roles may have vastly different levels of experience and knowledge of lung function.

Conflicting Opinions

As with most medical domains, it is an issue in lung function that many of the experts may have different opinions and define conflicting rules, depending on their own specialisations, teaching and experiences (a problem clearly shown in the results later presented in this study). These disagreements can be as simple as a

respiratory specialist having a more detailed understanding of some aspect of lung function than a GP; or two respiratory clinicians having encountered different borderline cases and forming different opinions on how to derive their results.

The effects of this are that any decisions or discovered data based on a single expert's opinions may not universally be considered correct: and so a widely applicable system, for either knowledge acquisition or knowledge discovery, needs to be able to adapt with conflicts and update its knowledge as it is in use.

2.5.2.2 Lung Function Data

Although the lungs perform very complex functions within the human body, they display few measurable outward signs of these functions (Laszlo, 1994). Even those indicators which are apparent are difficult to measure effectively, due to the execution of the test interfering with the normal process of breathing (Hughes & Empey, 1981; Ruppel, 1994). There are a wide range of tests used to gather data on functioning lungs, some more commonly used than others, some which are reliable only in certain situations, and most requiring specialist equipment.

The test results themselves are almost entirely comprised of numeric attributes, representing various measurements of the lungs and their functioning. Some other factors are also important for consideration, such as smoking history, sex, ethnicity; anecdotal evidence such as a medical practitioner's appraisal by sight and sound; and medical imaging techniques, such as the high resolution computed tomography (HRCT) tests used in the previously described ProcessNet study (Misra, et al., 2011); but all explicit lung function tests are measured as real numbers or ratios.

Lung Function Tests

No one test can provide a complete overview of all aspects of lung function (Hughes & Empey, 1981; A. Miller, 1987; Ruppel, 1994). Although each test measures different effects, using different means, all are essentially based on the same functions: this means that the information provided by these tests often overlap (Ruppel, 1994). This further means that combining the results of many of these tests can produce much more detailed information about the patient's lungs' function than would be available by a single test. Due to the uncertainty within any medical domain, caused by the incomplete understanding of medicine and the

complexities and wide degree of variation of the human body (Pribor, 1989; Tsumoto, 1998), any verification that can be provided from complimentary results from multiple tests will be beneficial in making conclusions with that data.

The lungs are one of the most difficult organs to measure without interfering with their function, as much of their function is involuntary: concentration is required to control our breathing, and it is usually such an unconscious action that our breathing patterns change significantly once we become aware of it and try to breathe normally[1]. The way we breathe is also affected by our emotional state, which can be influenced by the tests being performed. Other factors that can influence the measurements are the time of day, the temperature, and the state of the equipment used to take the measurements (M. Miller, et al., 2005). As such the measurements that are taken by these tests usually do not represent the normal function of the lungs but rather an approximation: observations on the normal breathing pattern of a patient, such as whether they are wheezing, are often reduced to being recorded anecdotally.

The tests themselves are generally divided into categories, based on which aspect of lung function is being measured and the inherent difficulty and cost in performing the tests. When referred to a specialist for lung function testing, not all patients will have all tests carried out: the more complex tests are only performed if the medical practitioner deems the basic test inconclusive and warranting further study (Pellegrino, et al., 2005).

Dynamic Spirometry

The most common series of tests are spirometry tests, as they are relatively inexpensive to perform and can be used as an indicator of some very common respiratory diseases such as Chronic Obstructive Pulmonary Disease (COPD) and asthma (Ferguson, Enright, Buist, & Higgins, 2000). These concern the volume change during specific breathing maneuvers (A. Miller, 1987), or in other words, the extent of the lungs' ability to move gas. Spirometry tests are undertaken with specialist equipment called a spirometer. To perform a test, the patient inhales as much as they can, then immediately exhales as fast as they can into a mouthpiece.

[1] As the reader may well attest

This procedure is repeated multiple times to get the best results, and for some key results, the highest measurements are taken even if they may be from different exhalations (M. Miller, et al., 2005; NACA, 2005).

Dynamic spirometry tests are measured relative to time, and provide a good example of how multiple measurements can be used together for more useful analysis. The measurement of the Forced Expiratory Volume over one second (FEV_1, or, the amount of air expelled in the first second of exhalation), while moderately useful on its own, is much more useful in conjunction with the measurement of Forced Vital Capacity (FVC, the total volume of air that can be exhaled from one breath). The FEV_1/FVC ratio, with consideration for certain other factors, is one of the most commonly used features in making key conclusions about a patient's lung function (Ruppel, 1994). Dynamic spirometry measurements are numeric values, recorded in litres or litres per second. In all there may be five or six different common spirometry measurements recorded from a test, with perhaps twice that many recorded in special circumstances (this is not including the further measurements derived by combining multiple other measurements, nor the repeat measurements made to ensure reliable results, nor the repeat measurements made after application of bronchodilator drugs).

Bronchodilator Response

A bronchodilator is anything which dilates the bronchi and bronchioles, the airways into the lungs. A bronchodilator drug is often administered during lung function testing to help determine where a patient's problem lies: a first run of spirometry test are performed, then a bronchodilator administered. Typically there is a short break to allow the drug time to take effect, then a second run of spirometry tests are performed (M. Miller, et al., 2005), although it has been recommended that for more reliable results the bronchodilators should be administered over time (Pellegrino, et al., 2005). Each measurement is then labelled pre-bronchodilator (pre-BD) or post-bronchodilator (post-BD). Typically, most measurements are retaken after a bronchodilator has been administered, effectively doubling the number of measurements. However the administering of bronchodilators is dependent on the purpose of the tests and the suspicions of the medical expert (M. Miller, et al., 2005).

Static Spirometry

The second category of tests is static spirometry, or lung volumes tests, which attempt to measure the full capacity of the lungs. They are made difficult because the lungs will always hold some gas that cannot be expelled, but this is taken into account with the procedures. Volume measurements are useful in identifying, clarifying, or eliminating many dysfunctions or problems, both new and previously identified by other tests (Laszlo, 1994; A. Miller, 1987). Lung volumes are numeric, usually recorded in litres. Approximately seven or eight different measurements of volume can be recorded in each session.

Diffusing Capacity

The final types of tests which make up the dataset are the diffusing capacity tests, which measure the ease with which gas is able to pass across the alveolar pulmonary cup membrane, thereby facilitating gas transfer. This category covers a very different aspect of the lungs to the other categories, and so can be used to identify specialised problem types. The measurements can also be used in combination with other results to identify complex problems, such as emphysema (Ruppel, 1994). The most common diffusing capacity test is the Diffusing Capacity of the Lung for Carbon Monoxide (D_LCO). It is measured by comparing the partial pressure difference between the inspired gas and the expired gas, the significant contributing factor to which is how effectively the lung can diffuse the carbon monoxide. This test has some ambiguity: as the carbon monoxide transfer is dependent on how much blood is present in the lung capillaries, anaemia can cause a reduction in D_LCO not directly related to the lungs. Because of this, many laboratories adjust the D_LCO based on the haemoglobin concentration found from a blood test, resulting in corrected and uncorrected D_LCO values (Crapo, Gardner, & Clausen, 1987). There are typically only three or four different measurements made in this category of tests.

Lung function tests generally do not provide enough information in themselves for a definitive diagnosis to be made; they will however provide insight into the nature of the patient's lungs and how they are functioning. In order to complete a diagnosis a medical expert will generally require much more detail: an examination of patient history, a physical examination, chest radiography, blood tests, sputum

examinations, and other tests, and a discussion with the patient themselves (Hughes & Empey, 1981).

Interpreting the Results

Interpreting these results is a complicated process that requires training and experience. While many medical practitioners will be able to identify and interpret basic conclusions using the simplest measurements, it is generally left to respiratory specialists to analyse and interpret the majority of lung function data The numbers of patients referred to lung function laboratories for assessment is relatively small, but the data produced is complex and only fully understood by a limited number of specialist clinicians.

Reference Equations

Lung function data can vary significantly depending on the ethnicity and living conditions of the patients being tested (Collen, Greenburg, Holley, King, & Hnatiuk, 2008; Subbarao, Lebecque, Corey, & Coates, 2004). It is therefore very important in analysing lung function data to consider these factors, and for this reason, as well as for simple practical reasons, studies are typically conducted with specific populations of subjects. It is also important as the results of one study cannot necessarily be generalised to any population of people.

Much of the complication in interpreting the data comes from this variability in results between people. As with any aspect of the human body, there can be large variability in the size and shape of the lungs and airways, and their function, depending on the subject; but most of the tests take absolute measures of capacity, speed, and function. The effect of this is that what may be a normal result for one person may be a critically bad result for another. This can be resolved by recording with the test measurements data such as age, height, weight, and sex, and taking these into consideration when interpreting the absolute measures of the other tests. Still, this is a difficult task, with potentially severe consequences for misinterpretation, so reference equations are used to determine expected or predicted values for each of the tests (Collen, et al., 2008).

The reference equations themselves are intermittently derived from large-scale studies into healthy populations. They can be derived from any study involving a

large population, such that extrapolating the trends in the population is reasonable. The equations are derived by sampling a large number of healthy patients through physical examinations, and using regression to define an equation incorporating factors such as age and height for each test in various demographic groups: typically sex and age but sometimes others (Crapo & Morris, 1981; H. Goldman & Becklake, 1959; Hankinson, Odencrantz, & Fedan, 1999; Subbarao, et al., 2004).

There are many reference equations to select from. However, as there are few large-scale studies performed, many are derived from specific populations. While they might make good predictors for those people, they are not necessarily reliable predictors for other populations. It is therefore important to choose equations that are based on a population as close as possible to the population under consideration (Collen, et al., 2008; M. Miller, et al., 2005; Subbarao, et al., 2004). Some of the more commonly used equations include those from the third National Health and Nutrition Examination Survey (NHANES III) study, which was a 6 year study carried out from 1988 to 1994 from a random sample of the U.S. population, collecting spirometry results from over 20,000 subjects (Hankinson, et al., 1999). Other commonly used equations come from studies by Crapo and Morris (Crapo & Morris, 1981; Crapo, Morris, & Gardner, 1981), Knudson (Knudson, Slatin, Lebowitz, & Burrows, 1976), and Quanjer (Quanjer, et al., 1993).

Limits of Normal

While there are equations that define the expected values for a patient's test results, using these for interpretation has become less commonly used, in favour of the more accurate approach of defining the upper and lower limits of normal for a patient. This method derives from the reference equations described above, by statistically calculating equations that determine what value should be considered above normal, and what value should be considered below normal. The limits of normal are commonly based on calculating the upper and lower 5^{th} percentile to derive the equations, providing a more accurate result than comparing a flat percentage of the predicted value (Pellegrino, et al., 2005). This approach is recommended by many leading respiratory bodies such as the American Thoracic Society (ATS) and European Respiratory Society (ERS) (Pellegrino, et al., 2005).

Difficulties with Interpretation

There are many factors influencing the difficulty involved in the interpretation of lung function tests, not the least being disagreements between experts over the best practices for doing so. Interpretation is also highly dependent on the geographic location and attributes of the population being tested (Collen, et al., 2008; Pellegrino, et al., 2005; Subbarao, et al., 2004).

Disagreement between experts

As with most medical fields, there can be significant disagreement between experts within the lung function domain over how to interpret results and what actions should be taken. A common source of disagreement is the differing nature of the professions that require lung function expertise. At a general level, the distinctions between clinicians and technicians often lead to disagreement. A common source is that they each have different goals when applying their expertise: a clinician's goal is to interpret the test results in an attempt to diagnose a patient's problem, in a limited timeframe, which necessitates a more practical perspective on interpreting the results. A technician's goal is to further develop the knowledge of how to perform and interpret the tests. Both can lend themselves to warranting greater accuracy or specificity, depending on the situation. A clinician will always want to be as certain as possible of their diagnosis to give their patient the best chance of recovery, without any risk of causing harm by misdiagnosis or failing to account for other conditions; but this often needs to be balanced with the expediency of treating a patient quickly. A technician however would be expected to be more concerned with making certain the results are as reliable as possible, with less pressure being applied on the timeliness of results. Additionally, clinicians generally have a much broader medical education. Their focus is on the diagnosis and treatment of a patient, a task in which lung function tests often have no relevance. Much of their understanding of finer details of lung function may therefore be learned through experience with patients, whereas a technician can, and is expected, to devote more time to published respiratory studies.

It is also unfortunately true that studies are not infallible and are rarely completely accurate: on occasions multiple studies have been published on the same topic, reaching different results. For example, there has been much discussion on how to

determine reversibility of airflow obstruction (Pellegrino, et al., 2005), with different studies reporting different results (Anthonisen & Wright, 1986; Eliasson & Degraff Jr, 1985). The GOLD controversy, described in more detail below, is another example. While these conflicts may be caused by mistakes in one or both studies, they are often caused by studies simply failing to take into account all the factors that may affect the results, in varying circumstances; which is unsurprising considering the vast expanse of data and possibility to be considered.

Human error

Human error is also an unfortunate factor that must be considered, particularly in clinical practice. Clinical experts are relied upon to remember everything they have learned, and to take into account every factor that might be important in interpreting the patient's test results, under constant pressure to return a diagnosis that they are confident in. Compounding this, they must commonly perform this for many completely different patients in a single day. With this expectation, it is inevitable that eventually even the most well-learned and precise clinician will miss something and make a misclassification or a misdiagnosis.

2.5.2.3 Standardisation of Knowledge

With the difficulties that can arise over how to best interpret lung function test results, it is unsurprising that attempts have been made to consolidate the experiences and knowledge of different professionals and develop standard sources of information for areas of lung function. These attempts are always long undertakings as they seek to, as much as is possible, complete the knowledge of a particular area; and they meet with differing levels of success. This section provides one example of an attempt at standardisation.

GOLD

The Global Initiative for Chronic Obstructive Lung Disease (GOLD) is an international organisation whose primary goal is to develop and maintain a global standard for diagnosing, managing and preventing Chronic Obstructive Pulmonary Disease (COPD) (GOLD, 2008). The first version of this report was published in 2001, having been written by an expert panel including "...a distinguished group of health professionals from the fields of respiratory medicine, epidemiology,

socioeconomics, public health, and health education" (GOLD, 2008), and the report has been updated many times since. Among many other achievements, this report describes a standardised approach to identifying and confirming a COPD diagnosis.

However, even after such a significant study there were disagreements with the published results as late as 2009. The GOLD report recommended the use of an FEV_1 value less than 80% of the predicted value, and an FEV_1 to FVC ratio of less than 0.7 to diagnose COPD. The lack of the lower limits of normal in this test caused many experts to question the accuracy of the report, saying that "GOLD has arbitrarily defined COPD on clinical and physiological criteria that have been argued to be not based on scientific evidence" (Kerstjens, 2004), and pointing out that this approach is not supported by the ATS or ERS. Studies were cited which showed that age affected lung volumes, and that consequently the use of predicted values for diagnosis would over-predict the disease in older people and under-predict it in younger people (Aggarwal, Gupta, Behera, & Jindal, 2006; Culver, 2006; Hardie, et al., 2002). Further studies were carried out to support this, and the results openly published in a fairly inflammatory style (Quanjer, 2009). In 2009 the GOLD report was amended to state "…because the process of aging does affect lung volumes the use of a fixed ratio may result in over diagnosis of COPD in the elderly, especially of mild disease. Using the lower limit of normal (LLN) values for FEV_1/FVC, that are based on the normal distribution and classify the bottom 5% of the healthy population as abnormal, is one way to minimize the potential misclassification", and that "…many experts recommend use of the lower limit of normal for each population" (GOLD, 2008).

This controversy highlights the potential for disagreement between experts in the domain: even creating a standardised approach to identifying a single diagnosis from minimal attributes can lead to disagreements which can require some time to resolve. Attempting to find a balance between ease of use, applicability to a wide population, and accuracy of results can be a difficult task, even without considering any errors or statistical anomalies.

2.5.2.4 Lung Function Computational Studies

As with most major medical fields, lung function has been the subject for expert system development and knowledge discovery. The most well known of these is the

PUFF expert system for interpreting lung function data, but this has been followed by other studies and developments.

Expert Systems

While PUFF is the most well known expert system for lung function, at least in research literature, there have been more recent systems developed. One example is Pulmonary Consult, a commercial product from the Medical Graphics Corporation (MedGraphics, 2011). As it is a commercial product little detail is available on its development and content; however it is known to have been built upon the knowledge base from PUFF and so largely covers the same area (Thomson, 2009).

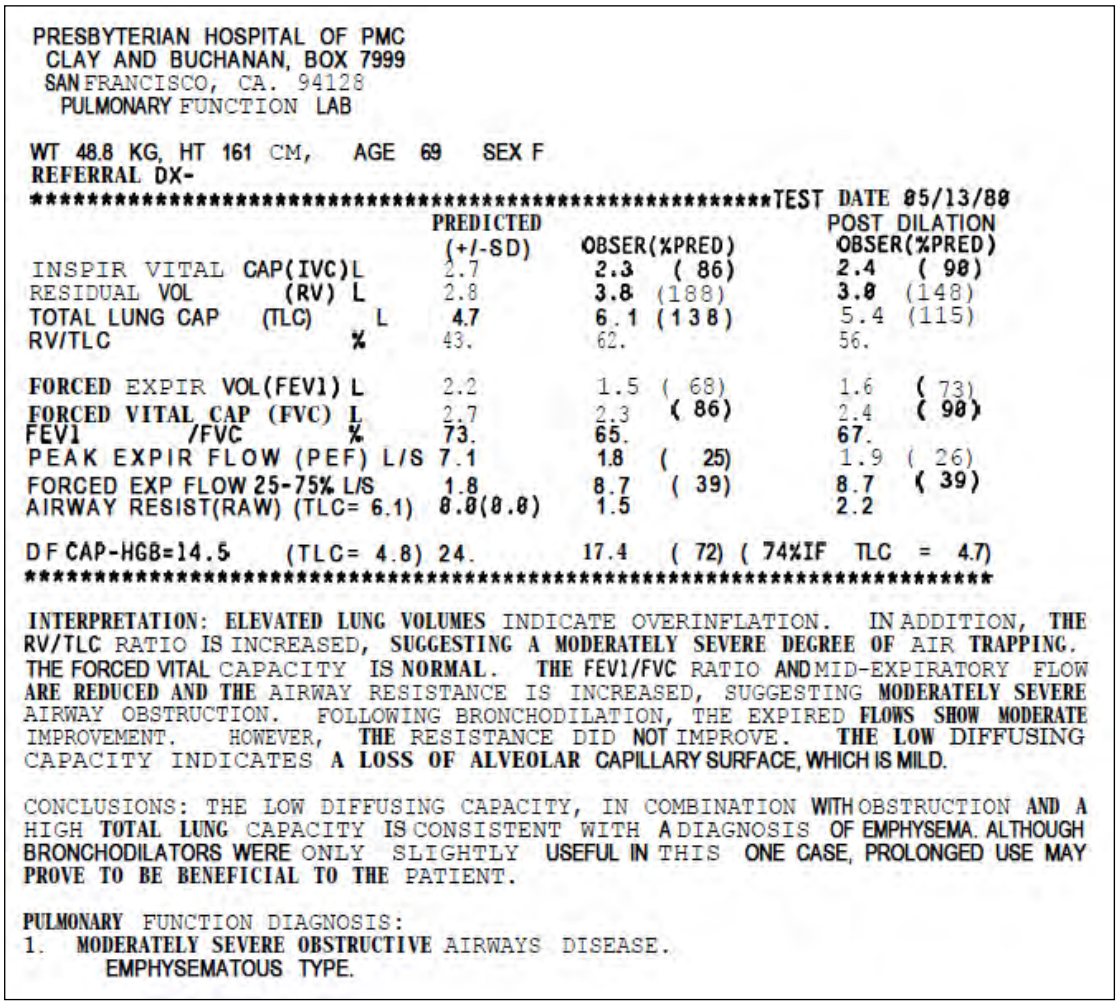
PUFF

PUFF was developed in the early 1980s as a test of the Essential MYCIN (EMYCIN) framework, which was a generalisation of the MYCIN expert system such that it could be applied to different domains. PUFF was deployed in the Pacific Medical Centre in San Francisco to assist pulmonary physiologists in interpreting the results of patient lung function tests, by taking in spirometry, lung volume, and diffusing capacity test results and returning interpretations based on the rules in its knowledge base.

The reasons for the expert system being developed in the lung function field were many: the interpretation of lung function tests is a daily problem, and so fills a need; the interpretation task was complex enough to be challenging; the lung function data was mostly self-contained, not requiring large amounts of data apart from that gathered in the lung function tests; there was available data; expert interpretations tended to be phrased similarly; and there was significant tedious work involved for the experts in generating reports.

It used classification rules, an inference engine, a knowledge acquisition module, and an explanation module. The system would function by asking the user, a lung function expert, a series of questions about the current case, thereby building the data about the current case. Once received, it would infer from that data and the rules in its knowledge base interpretations, which it would respond with. Over 4 years, the system interpreted over 4000 cases, providing interpretations for approximately 10 patients each day in use. Figure 2-8: shows a sample of the output

from PUFF's interpretation, following a standard lung function report format where possible.



PRESBYTERIAN HOSPITAL OF PMC
CLAY AND BUCHANAN, BOX 7999
SAN FRANCISCO, CA. 94128
PULMONARY FUNCTION LAB

WT 48.8 KG, HT 161 CM, AGE 69 SEX F
REFERRAL DX-
**********TEST DATE 05/13/80

	PREDICTED (+/-SD)	OBSER(%PRED)	POST DILATION OBSER(%PRED)
INSPIR VITAL CAP(IVC)L	2.7	2.3 (86)	2.4 (90)
RESIDUAL VOL (RV) L	2.8	3.8 (188)	3.0 (148)
TOTAL LUNG CAP (TLC) L	4.7	6.1 (138)	5.4 (115)
RV/TLC %	43.	62.	56.
FORCED EXPIR VOL(FEV1) L	2.2	1.5 (68)	1.6 (73)
FORCED VITAL CAP (FVC) L	2.7	2.3 (86)	2.4 (90)
FEV1 /FVC %	73.	65.	67.
PEAK EXPIR FLOW (PEF) L/S	7.1	1.8 (25)	1.9 (26)
FORCED EXP FLOW 25-75% L/S	1.8	8.7 (39)	8.7 (39)
AIRWAY RESIST(RAW) (TLC= 6.1)	0.0(0.0)	1.5	2.2
D F CAP-HGB=14.5 (TLC= 4.8)	24.	17.4 (72)	(74%IF TLC = 4.7)

****

INTERPRETATION: ELEVATED LUNG VOLUMES INDICATE OVERINFLATION. IN ADDITION, THE RV/TLC RATIO IS INCREASED, SUGGESTING A MODERATELY SEVERE DEGREE OF AIR TRAPPING. THE FORCED VITAL CAPACITY IS NORMAL. THE FEV1/FVC RATIO AND MID-EXPIRATORY FLOW ARE REDUCED AND THE AIRWAY RESISTANCE IS INCREASED, SUGGESTING MODERATELY SEVERE AIRWAY OBSTRUCTION. FOLLOWING BRONCHODILATION, THE EXPIRED FLOWS SHOW MODERATE IMPROVEMENT. HOWEVER, THE RESISTANCE DID NOT IMPROVE. THE LOW DIFFUSING CAPACITY INDICATES A LOSS OF ALVEOLAR CAPILLARY SURFACE, WHICH IS MILD.

CONCLUSIONS: THE LOW DIFFUSING CAPACITY, IN COMBINATION WITH OBSTRUCTION AND A HIGH TOTAL LUNG CAPACITY IS CONSISTENT WITH A DIAGNOSIS OF EMPHYSEMA. ALTHOUGH BRONCHODILATORS WERE ONLY SLIGHTLY USEFUL IN THIS ONE CASE, PROLONGED USE MAY PROVE TO BE BENEFICIAL TO THE PATIENT.

PULMONARY FUNCTION DIAGNOSIS:
1. MODERATELY SEVERE OBSTRUCTIVE AIRWAYS DISEASE.
EMPHYSEMATOUS TYPE.

Figure 2-8: PUFF sample report output (Aikins, et al., 1983)

It was concluded that PUFF was a ~~"p~~ractical assistant to the pulmonary physiologist" (Aikins, et al., 1983), as it had the support of hospital staff and administration and was in daily use. However, areas for improvement were noted. The system lacked the ability to identify prototypical patterns; there was difficulty involved in adding new knowledge to the system, as the addition of a new rule may affect the behaviour of existing rules in unexpected ways; there were problems with the order that data was requested; and it lacked the ability to adequately explain the results that were reached (Aikins, et al., 1983).

Pulmonary Consult

Pulmonary Consult is another expert system for assisting in the interpretation of lung function test results. It is a commercial product from the Medical Graphics Corporation (MedGraphics, 2011), and as such little detail is available on its development and content; however it is known to have been built upon the knowledge base from PUFF and so largely covers the same area (Thomson, 2009). It has been available since the 1980s and is used in many clinical settings.

Knowledge Discovery

There have been surprisingly few applications of data mining and knowledge discovery to the field of lung function. Numerous studies have been performed in highly specialised areas of lung function, such as analysing a thoracic lung cancer database (J. Goldman, Chu, Parker, & Goldman, 2008), an association study between gene variations and bronchopulmonary dysplasia attempting to find the causes of that one lung disease (Rova, et al., 2004), and another data mining study into a dataset of a specific lung abnormality (solitary pulmonary nodules) (Kusiak, Kern, Kernstine, & Tseng, 2002). Other studies have also been performed on data related to lung function, such as a case based reasoning approach to automatically building a classifier for molecular biology, which also tested over a lung microarray dataset (Arshadi & Jurisica, 2005). However, there has been very little work into broader attempts to analyse lung function test data, and almost no exploratory data mining: all data mining studies in lung function seem to be explanatory in nature, trying to find detailed reasons for specific events or phenomenon.

Exposed MCRDR

An approach which combined MCRDR knowledge acquisition, data mining and expert-driven analysis was developed in 2006 (Ling, 2006). The method, given the name Exposed Multiple Classification Ripple-Down Rules (EMCRDR), was based on the premise that the MCRDR methodology would allow the acquisition of a strong knowledge base. From that base, experimental hypotheses could be added as new ―knowledge", which would then be validated (or not) through the MCRDR validation process: allowing an exploratory approach to knowledge discovery. The validation process would use a dataset to provide evidence for the hypothetical knowledge, point out the inconsistencies, and assist in developing the hypothesis

until it was compatible with existing knowledge and data. It also suggested that extra validation mechanisms might be added to allow the expert to further verify that their hypotheses were supported by the data, and a rudimentary data mining feature was added that could either assist in defining rule conditions to match a group of cases, or could identify the cases that matched conditions defined by the expert.

Modifications to the MCRDR process

The study contained a few small but significant modifications to the basic MCRDR approach to facilitate the knowledge discovery application. The most significant of these modifications was to allow the expert free access to view and modify the knowledge base, to the extent of being able to edit or delete existing rules. This is in direct contrast to the traditionally accepted wisdom in RDR development that the knowledge base only ever be added to, never edited or deleted from (Compton & Edwards, 1994). Exception rules and stopping rules provide all the functionality of editing and removing without invalidating the context of any existing knowledge (Kang, 1996).

The second significant departure from a normal MCRDR implementation was the addition of a dataset, which caused the cornerstone case model to be much different: rather than the cornerstone cases being any previously seen case which matched a rule when the rule was made, the EMCRDR system maintained a list of all classifications for all cases in the dataset. When the expert was defining a rule, all cases matching the rule would be displayed, and this was used to provide validation for the rule.

Impact of EMCRDR modifications

The EMCRDR study tested a small dataset of approximately 400 cases, with one expert, in the domain of lung function. While it found evidence to suggest that the EMCRDR approach worked, the study was far from conclusive (Ling, 2006). It also made no conclusions as to how well the method worked. It did however highlight many areas for potential improvement; in particular, the study demonstrated the effects of the modifications to the base MCRDR approach and how they might be better adapted and applied.

It was found that while allowing the expert to view the knowledge base provided a relatively effortless way of expressing the existing knowledge, it caused a shift of focus from a case-based expression of knowledge to a rule-based one; and unfortunately this shift invalidates many of the advantages of the MCRDR approach. It requires the expert to understand precisely how the knowledge base works in order to add rules correctly, which is an unrealistic expectation. Given the inevitable restrictions on expert time mentioned previously, it is in most instances impractical, if not impossible, to take the time to teach the expert exactly how the rules they define inter-relate. Depending on how the knowledge base was built, understanding exactly how the rules combine and what applies in any given instance can be a very difficult task regardless of how familiar the person is with the MCRDR method. Supporting this, the study reported that the expert struggled with determining exactly what rules should be applied where to achieve the desired results (Ling, 2006). This was attributed to the contrast between the traditional MCRDR implementation style which hides the structure of the knowledge base, and the more explicit style used in places to show the structure of the knowledge base. The combination of these conflicting styles and the inherent problems with the expert understanding how the knowledge base works resulted in confusion from the expert on how to add rules, which type of rule to use, and how to solve errors in the knowledge base.

As a direct consequence of this confusion, the ability to edit and delete rules was very rarely utilised, and mostly to little effect. It was noted that the expert liked having the ability to edit rules and used it commonly to correct small mistakes; it is suggested however that had the expert had a better understanding of how to define rules to begin with, less errors would have required correction. Rule deletions were very rarely used, and seemed to offer no real benefit over the normal stopping rules, and in fact may have hindered progress as at least stopping rules could have provided an indication of the problems the expert was encountering.

A related issue, not discussed in the study but possibly the underlying cause of much of the confusion, is that when the knowledge base is viewed as a single entity it is missing the integral components of context and evidence. Although a parent rule gives an outline of the context for its exception rules, without the context of the cases themselves there can be significant missing information. It has been noted by

many that the knowledge added to a knowledge base, in any form, is not concrete: it may (and is even likely to) change over time, and it may change when presented with a different context of application (Compton & Jansen, 1989). This means that when a rule is considered outside of the context it was made in, and without a framework of data showing how it is applied, there is a stronger chance that the expert will misunderstand the intention and application of the rule. Similarly, while the method in the study provided a list of all the cases that the new version of a rule covered, it gave no indication of which cases were no longer covered by the rule. As such, a case which the expert had previously decided was complete, and hence would be very unlikely to look at again, could now have different results. This again reinforced the rule-centred mode of thinking which was determined to be detrimental to the knowledge acquisition and discovery process.

A further problem raised in the study is that, because the expert is no longer considering individual cases until they are completed, it would be expected that the ability to derive tacit knowledge is reduced, or even removed completely (Ling, 2006). Traditionally an expert will consider one case at a time and continue working with that case until they are satisfied that it is completely correctly classified. However under a rule-focused approach, the expert will define rules for their most commonly used knowledge first, in the order that the knowledge occurs to them, without completing their current case. Unless they revert to a case-focused approach later, they will likely miss some of the rarer conclusions as the expert is unlikely to recall them from memory without prompting. Tacit knowledge – that is, knowledge which is difficult to define and express – will likely be missed completely as the expert is not given a situation requiring such knowledge.

This also has serious implications for the validation process. If there are no cases completely reviewed then indications of cornerstone case conflicts will be less likely and less meaningful. This problem was addressed in the EMCRDR study by the removal of traditional cornerstone case validation, and instead while a rule was being created, showing all the cases in the dataset that match the rule. The expert would then look through each of those cases to determine if the rule is correct. This is clearly an impractical solution for any sizable dataset, requiring the expert to look through and classify potentially hundreds, even thousands, of cases to check that each rule is correct. Also, for this validation strategy to be effective at all it requires

that the dataset be representative of the frequency and range of cases in the domain – and the typical way to ensure an unclassified dataset has these attributes is to use as large a dataset as possible. While a size balance might well be found between the two, it is a fundamental issue that any method which does not take advantage of all the data available will not be as effective as it could be. Also, the lack of cornerstone validation meant that the expert was required to examine and evaluate every case in the dataset which matched their new rule, in order to find those that invalidated the rule or provided additional information, if any existed.

However, the approach was found to achieve the desired goal, with the expert discovering new knowledge and being apparently satisfied with the method and the result. It did however highlight many areas for potential improvement, particularly in resolving the issues with rule-based thinking and errors in knowledge acquisition, misunderstanding the knowledge base, and the potential for providing data mining assistance to the user.

2.6 Summary

The literature described here shows that knowledge discovery is a complex, multi-stage process (Fayyad, et al., 1996a; Kurgan & Musilek, 2006). Of these stages, the data analysis or data mining stage has been heavily researched, with many methods available for finding patterns in data (Brachman & Anand, 1996; Witten & Frank, 2005). However, these methods encounter difficulty in analysing complex data, particularly when there is a large or complex existing body of knowledge about the meaning of that data (Liu, et al., 1997; Piatetsky-Shapiro, et al., 1994; Sinha & Zhao, 2008). Under various knowledge discovery models, this problem is addressed by the identification and incorporation of knowledge in the initial stages (Fayyad, et al., 1996a; Kurgan & Musilek, 2006): however, it is an identified problem that there few methods have been developed to achieve this (Sinha & Zhao, 2008), particularly for domains or applications of a realistic complexity (Adejuwon & Mosavi, 2010; C. Zhang, et al., 2009). The purpose of this study is therefore to develop and test a new method that can effectively acquire and incorporate existing knowledge into a knowledge discovery process, for a complex domain. As the literature has shown, the lung function domain is suitably complex and can benefit from such data analysis (Cios & Moore, 2002a; Roddick, et al., 2003).

However, in developing this method a number of issues are raised. The first component is to acquire the knowledge to be incorporated: but as this chapter has shown, there are many methods available for acquiring knowledge from data. Of those methods that have been tested in assisting knowledge discovery, a commonly identified problem is that they require an impractically large knowledge acquisition or knowledge engineering commitment to find effective results (Kotsifakos, et al., 2008; Liu, et al., 1997; C. Zhang, et al., 2009). Another important concern is that the knowledge acquisition process must be able to be updated incrementally, as the knowledge required will change over time (Liu, et al., 1997; Piatetsky-Shapiro, et al., 1994). MCRDR is an incremental knowledge acquisition method that has been shown to help resolve problems with knowledge acquisition and engineering requirements (Kang, 1996; Kang, et al., 1995), and so this seems a logical choice to build the required knowledge base. Chapter 3 presents the results of acquiring a knowledge base from lung function experts, including the impact of using a MCRDR process modified to take advantage of the availability of a dataset.

The next question raised in this study is how the acquired knowledge base can be applied to a knowledge discovery task. A method was developed to achieve this based on the knowledge acquisition framework; this method and its efficacy at discovering new lung function knowledge is tested in Chapter 4 of this thesis.

The final issue raised is that conflicts in knowledge can occur in the lung function field, with different experts reaching different conclusions and having different understandings of the data (Pellegrino, et al., 2005; Quanjer, 2009). To help resolve this issue in the domain, and to ensure that the acquired knowledge was as accurate as possible, a method was needed to compare and assist in the consolidation of the knowledge of multiple experts. Existing methods for comparing or integrating MCRDR knowledge bases lack a focus on improving the knowledge of the experts involved, and do not take advantage of available data (Beydoun, et al., 2005; Richards, 2009; Vazey & Richards, 2006). Therefore a method was developed to quantify the differences between acquired knowledge bases, and provide evidence to assist in resolving conflicts. In addition to being tested with the knowledge bases acquired from the experts, this method was tested by comparing acquired student knowledge with the combined expert knowledge, as a potential teaching and

assessment tool. The method used and the results of both these tests are presented in Chapter 5.

Chapter 3 An Expert System for Lung Function Interpretation

3.1 Introduction

When attempting to discover new knowledge by analysis of data, the chance and magnitude of success can be greatly improved by the establishment of a layer of initial knowledge, adding meaning to the data and guidance to the analysis (Fayyad, et al., 1996b). This is particularly true of complex data with large quantities of existing knowledge. This knowledge can vary greatly in complexity: ranging from an understanding of the relevance of attributes, and assigning them weights based on significance; to having the ability to identify patterns and groupings within the data, and derive some new information implied by that pattern. Identifying and using this knowledge effectively can dramatically increase the efficacy of the analysis and discovery, by both guiding how to analyse the data and in helping to determine the usefulness of the result (Fayyad, et al., 1996b; Pohle, 2003).

However there are further benefits to the acquisition of such knowledge. Once acquired in a computer-useable form, that knowledge can be applied as an expert system, capable of automatically providing an expert analysis of similar data. This can be desirable for a number of reasons, for example: if expertise is scarce, training another expert is usually helpful; expert systems can be duplicated and so expertise can be spread; computers can in many areas be more reliable than human experts, improving the accuracy of the task being performed; common tasks can be automated allowing the human experts to work on higher-level challenges; and the knowledge can be put to use in building more advanced systems.

In the domain considered in this study, many of these problems are evident. While a low level of expertise is common, high level expertise is quite scarce. Reliability and consistency can be a problem. Common tasks do take up a significant portion of the experts' time. A further challenge is that because low-level knowledge is widespread but high-level knowledge is rare, and much of their expertise is learned tacitly, experts' opinions can differ greatly concerning how to interpret the data and

what actions should be taken in certain situations. Each of these factors makes an effective expert system a desirable tool.

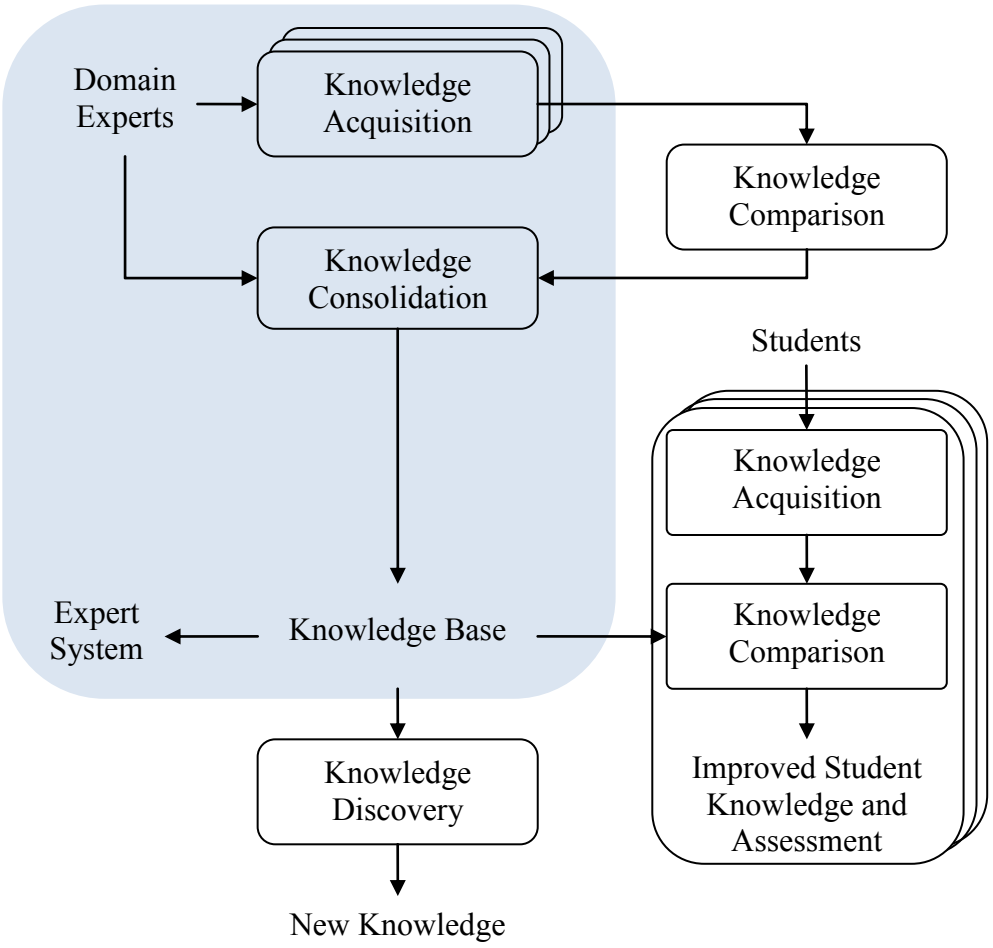


Figure 3-1: The methods presented in this thesis; the highlighted section shows the components presented in Chapter 3

This chapter presents the results of developing a knowledge base and expert system for the interpretation of lung function data, using Multiple Classification Ripple Down Rules (MCRDR). With the availability of a large dataset the opportunity was taken to implement additional data-based validation, to improve the efficacy of the knowledge acquisition. Work was also undertaken to acquire and consolidate the knowledge of multiple experts, using methods for collaborative development and the integration of multiple knowledge bases. The effects of these modifications on the knowledge acquisition process and their potential for future development are discussed. The developed knowledge base provides the capability of an expert system, and is integral to the methods presented in Chapter 4 and Chapter 5: by

assisting in data analysis for knowledge discovery, and by providing a standard of knowledge to compare against student input. Figure 3-1 shows these components in the context of the larger study.

3.2 Methodology

This section describes the modified MCRDR method used to develop a knowledge base for decision support in the interpretation of lung function data, and for use in assisting knowledge discovery. This method was implemented as an online system for examining the data and entering knowledge. In acquiring this knowledge, particular datasets and human experts were available which influenced the nature of the knowledge acquisition process. These resources will be described first, followed by an explanation of the impact of those resources on the knowledge acquisition design.

3.2.1 Lung Function Resources

The availability of both data and experts had a large influence on the direction of this study. The study was prompted in part by the availability of large numbers of archived lung function reports, a resource with the potential for expanding current knowledge. Whilst a lack of expert availability also shaped the course of the study to some extent, this resource deficiency highlights the potential benefits of research in this area.

3.2.1.1 Data

The data that was used to acquire the domain knowledge consisted of an amalgamation of lung function case reports from three sources: 1568 reports from Austin Health in Melbourne, Australia[2]; 1390 reports from the 2004 round of the Tasmanian Longitudinal Health Study (TAHS)[3]; and 5 reports from the Royal Hobart Hospital in Hobart, Australia[4]. Each report was considered to be a single case in the dataset, with the source added as a further attribute. In the implementation of the knowledge acquisition system, each of these sources were

[2] http://www.austin.org.au/

[3] http://www.epi.unimelb.edu.au/research/major/tahs

[4] http://www.dhhs.tas.gov.au/hospital/royal-hobart-hospital

presented as distinct datasets, but also with the option to view them all as a single dataset; as it is common for experts to be interested only in a single source of data for knowledge discovery purposes. The data was considered similar enough that they would be unlikely to want to define a rule for only one dataset, but the inclusion of the source as an attribute allowed this if necessary.

All cases had any identifying data removed for privacy reasons and were identified within the online system by an ID number and Source pair; for example, ~~"c~~ase 38 from the TAHS study", or ~~"3~~8 TAHS". The Source was used as an identifier to allow for the possibility that cases may be linked back to the archived stores, which may have additional information for future analysis.

Importantly, all cases in the dataset were entirely unclassified – they had no information such as eventual interpretations or diagnoses, nor any information on the future discovered effects for each patient. This precluded automated machine learning approaches from consideration in developing a knowledge base for the data. Each case constituted a single set of test results from a single patient, independent of history, future tests or information, or any form of information other than the recorded test results.

When presented to the users, all reports were displayed in a format similar to the printed formats that are used by most medical institutions, so as to be recognisable and familiar. Figure 3-2 shows an example lung function report, as they appear in the online system.

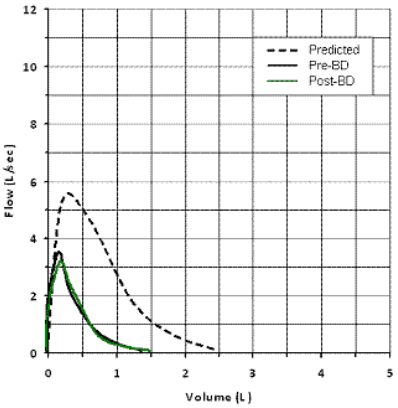
As Figure 3-2 shows, not all cases had values for all attributes, and many were missing values for different attributes. The exact measurements taken may have depended on the facility where the tests were performed, the reason the tests were being performed, who was performing the tests, practical restrictions due to other medical problems, or even broken equipment – the reasons for their omission were not recorded with the cases. These missing values had little impact on the knowledge acquisition process, as most cases contained sufficient information for classifications to be made; and if any single case did not, the definition for classifications could be derived from other cases.

Case 87 from AH View Conclusions Switch to Tab Report

Patient ID:	87	Age:	57	Sex:	F	Date:	2000	Weight:	80
Test ID:	87	Smoker:		Pack Years:	0	Ethnicity:	Cauc	Height:	150	BMI:	35.56

	Pre - Bronchodilator			Post - Bronchodilator		
	Actual	Predicted	% Predicted	Actual	% Predicted	% Change
Spirometry						
FEV1 (L)	0.91	2.08	43.8	0.92	44.2	1.1
FVC (L)	1.36	2.51	54.2	1.47	58.6	8.1
FEV1/FVC	0.67	0.82	81.7	0.63	76.8	
PEF (L/sec)	3.11	5.19	59.9	2.8	54.0	
FEF 25% (L/sec)	1.98	4.89	40.5	2.27	46.4	5.9
FEF 75% (L/sec)	0.25	1.08	23.2	0.12	11.1	-12.0
FEF 25-75% (L/sec)	0.52	2.30	22.6	0.38	16.5	-6.1
Lung Volumes						
SVC (L)	1.43	2.51	57.0			
IC (L)	1.14	1.80	63.3			
ERV (L)	0.29	0.71	40.9			
RV (L)	1.67	1.68	99.4			
TLC (L)	3.09	4.19	73.8			
RV/TLC (%)	0.54	0.38	142.1			
FRC (L)	1.96	2.39	82.0			
Diffusion						
DLCOcor (ml/min/mmHg)	10.09	19.98	50.5			
DLCOuncor (ml/min/mmHg)	10.2	19.98	51.1			
VA (L)	2.01	4.00	50.3			
TLC - VA (L)	1.08					
DLCOunc/VA (ml/min/mmHg/L)	5.07	5.00	101.4			
DLCOcor/VA (ml/min/mmHg/L)	5.02	5.00	100.4			
Blood Gases						
Hgb	15					

Flow Volume Loop



Close Report Back to Main Window

Figure 3-2: Sample lung function report

Reference Equations

As described in section 2.5.2, lung function reports are typically presented with the predicted values for each attribute, as determined by a set of reference equations. In this study, the report data initially contained values derived from the Knudson et al. 1976 equations (Knudson, et al., 1976) for spirometry, gas transfer equations from Cotes and Leathart (Cotes & Leathart, 1993), and Goldman and Becklake's 1959 lung volumes equations (H. Goldman & Becklake, 1959). During the developmental stages of this study however, these were rejected by a number of lung function experts as being somewhat outdated, and new equations were introduced: the NHANES III equations for spirometry (Hankinson, et al., 1999), the Roca et al. equations for lung volumes (Roca, et al., 1990), and the Quanjer et al. equations for gas transfer (Quanjer, et al., 1993). The previous equations were kept as an option to allow experts to compare results between reference equations, and to allow them to use whichever they felt was most appropriate.

Flow-Volume Loops and Data Visualisations

There were however a few significant differences between the reports that experts typically see in their professional work and the reports that were presented with the system. Significantly, most cases in the system did not have the Flow-Volume Loop (FVL) diagrams, volume graphs, or any associated visual representation of the test results. FVL diagrams visually describe the airflow during the inhalation and exhalation measured in spirometry. This is generally considered to be a vital component in the interpretation of a lung function report, as the visual cues provided by the shape of the FVL provide respiratory experts with an immediate impression of what to be looking for and how to proceed, and often an initial diagnosis. It is also considered to be critical both by inexperienced experts who are not as aware of the significance of all of the attributes, and by experienced experts who can infer a great deal from an initial glance. While the FVL generally does not provide any information that the test results do not, it has become such an effective shortcut to interpreting results that it is expected on reports and some experts come to rely upon it for their interpretations. In fact, when experts were initially approached to take part in this study many were uncomfortable working without FVL and declined to take part (this appeared to be entirely based on personal preference, with the type and experience of experts not providing any indicator of whether they would refuse). As they can have such a critical role, reports were added to the dataset from the Royal Hobart Hospital with the FVL and volumes graphs attached; and twenty more FVL were created by a leading respiratory scientist to match a set of cases chosen to be representative of the range of cases in the dataset, to allow as many experts as possible the opportunity to participate in the study.

3.2.1.2 The Experts

In an effort to ensure the best possible resultant knowledge base, multiple experts were used to perform the knowledge acquisition. These experts had a range of experience and knowledge in the lung function field, in working with patients and performing respiratory research.

Three experts were used to acquire the knowledge for the main knowledge base in this study. Primarily the knowledge came from a single leading respiratory scientist

in Australia, with additional input from another highly regarded clinical specialist, and some minor additions by another respiratory researcher. Further input, ranging from system design and testing to explaining complexities of the lung function domain, was taken from 15 more available experts in Australia.

Initially, in order to organise his thoughts and establish some fundamental classifications, the leading expert created a document detailing definitions for a set of common classifications. This document was circulated and confirmed by another small group of respiratory experts, including the two other experts involved in developing the knowledge base. Once confirmed, the administrator of the system added these definitions to the system as a basic set of initial rules, in the manner of a Vazey CARD approach (Vazey, 2006). The secondary expert then contributed to this knowledge base, along with the tertiary expert. As the experts were not concurrently available, and in order to allow knowledge comparisons, the first expert also developed their own knowledge base independently of the collaborative knowledge base. Finally, the two knowledge bases were compared, inconsistencies were resolved where necessary, and the knowledge bases consolidated into a single final knowledge base. The development and contributions towards each knowledge base are summarised in Figure 3-3. The methods for acquiring and consolidating the experts' knowledge are discussed in the following section.

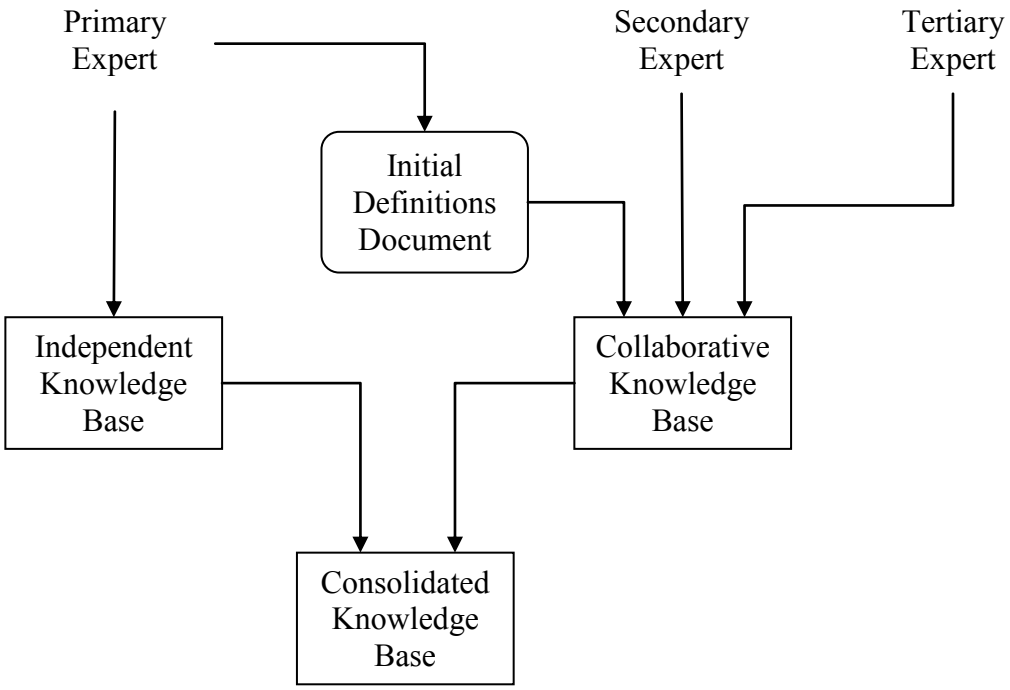


Figure 3-3: Contributors to each knowledge base

3.2.2 MCRDR Implementation

For the most part, the implementation of the MCRDR knowledge acquisition system was as per the original MCRDR implementation by Kang (Kang, 1996), with some functional limitations and modifications necessitated by both the specifics of the domain, and the dual application as a knowledge discovery device.

3.2.2.1 Standard Features

Rules

Rules in general were handled as per a typical MCRDR system: experts could define root level, exception, and stopping rules. Again as with most implementations, each rule condition could only be conjuncted, with no option for disjuncts, and hence no need for condition grouping within a rule (such as [(condition1 AND condition2) OR condition3)]). Although MCRDR can use disjuncts perfectly well, this was decided in an effort to keep rule creation standardised and to make comparisons between different experts' rules as simple as possible. If ever necessary, disjuncts could be handled by adding multiple rules reaching the same conclusion.

Conditions

Rule conditions themselves were kept in a very simple format: `[Case attribute] [Operator] [Value]`, where `[Case attribute]` is a field name for an attribute, operator is from the set {< (is less than), <= (is less or equal to), > (is greater than), >= (is greater or equal to), = (is equal to), != (is not equal to), missing (value is unknown), not missing (value is known)}. This is more restrictive than many rule-based systems which allow a condition to directly compare two attributes (for example `[attribute1] [is less than] [attribute2]`). This method was partially chosen for simplicity, but also as it parallels the traditional form used within the lung function domain – if making a decision based on two attributes of a case, the typical approach is to derive a new attribute from some calculation incorporating those two, and refer to that derived value: for example, the ~~F~~EV_1 to

FVC ratio" is often discussed, and is always included in reports as a distinct attribute (Pierce, et al., 2005; Subbarao, et al., 2004).

Classifications

In order to allow the experts as much freedom of expression as possible rule classifications could be defined as free-form text. When defining a rule the expert was presented with a list of all classifications currently entered into the system from which to choose; and if the exact classification they desired was not present, they were given the option of defining a new classification. Each classification consisted of a classification title of up to fifty characters, intended to be a short phrase or description, and an optional accompanying statement of up to 255 characters, intended to be a short paragraph or two explaining anything that may be ambiguous about the classification or any further detail the expert felt needed to be expressed.

3.2.2.2 Novel Features

The application as a data mining tool necessitated a number of modifications to the implementation of the MCRDR system, and provided an opportunity for others. The change with perhaps the most impact was shifting from considering individual cases as they are presented, to considering a dataset as a whole. This relatively minor difference resulted in a cascade of other modifications and adjustments. The changes made to the basic MCRDR process were:

- Presenting a dataset, not a case
- Defining potential cornerstone cases based on expert acceptance
- Dataset statistics for additional validation
- Working with multiple experts: collaboration and consolidation

Each of these modifications are described in this section.

MCRDR with a dataset

The first difference to traditional MCRDR knowledge acquisition was that rather than cases being serially presented to the expert, a large number of cases were displayed as a set. This approach was necessary to facilitate the knowledge discovery component of this study: when trying to develop new domain knowledge, the expert needs to be able to examine cases matching specific criteria that they are

interested in, and to request summary information about sets of cases. For example, the expert may wish to compare the average age of different classification categories; or they may wish to know the proportion of cases with one classification which also have a second classification. In general, in order to perform any analysis of the lung function domain, or a subset of the domain, the user must be able to consider a dataset holistically and not as individual, disconnected cases. Given the interconnected nature of the knowledge discovery and knowledge acquisition components of this study (described in more detail in Chapter 4), and that the experts must be aware of all the components, the knowledge acquisition was by necessity performed with a dataset rather than through individual cases.

Performing knowledge acquisition from a set of cases, rather than strictly incrementally with cases ―as they occur", effects the implementation in many subtle ways. Whereas a typical MCRDR system might involve ―loading a case" as the first step of any interaction, this implementation requires that the expert load a dataset, and choose a case to work with. From this point, the specifics of the knowledge acquisition are identical: the expert works through that individual case, then works on another single case. However, the subtle distinction in selecting a case from a set, rather than being presented with an individual case, can result in the previously mentioned problem of a rule-centred, rather than a case-centred, approach to knowledge acquisition.

There are however benefits to presenting the user with a dataset. As noted by Vazey's CARD approach, some rule-centred thinking helps the user to establish their knowledge base (Vazey, 2006). In addition, using a dataset provides a representation of the domain, and as such provides the potential for better validation in the rule making process. The methods used to achieve this are discussed further in the following sections on cornerstone cases and statistics.

A functional change to knowledge maintenance was also made. In keeping with the incremental nature of the MCRDR process, whenever a new rule is added to the system all cases in the dataset are checked to see if they match. Records are then created linking each case, classification, and the rule that caused it. These records were created and maintained for efficiency purposes, as many of the further interactions with the dataset such as displaying statistical information, manipulating the dataset, and providing feedback for validation are made much faster by having

this information stored rather than repeatedly inferring it through the knowledge base. Correctness is maintained by updating these records as necessary whenever any change is made to the knowledge base.

Cornerstone cases

The inclusion of a dataset also presents an opportunity to improve the effectiveness of cornerstone cases. The traditional cornerstone case mechanism uses cases that have been previously seen to validate any new rules that are entered, by ensuring that a new rule does not change the results for a case that has already been seen. If we are working with a set of cases however, the system has effectively ―seen" all of them already.

In place of cornerstone cases, the EMCRDR study took the simplistic approach of showing all cases that matched the current rule (Ling, 2006). This was useful to an extent, in that it provided some feedback on what effect the rule would have on classifying the dataset; furthermore this feedback is above what is normally provided by cornerstone cases, as it shows the effect of the rule over a wider coverage of possibilities. However, it only provided a benefit when defining quite specific rules, and when using a minimal dataset, as the sheer number of cases presented could be completely impractical to review. Furthermore, the expert would only receive indications of conflicts with previous decisions if they reviewed and classified each case individually, making this a poor substitute for cornerstone case validation.

In this study a combination of a traditional cornerstone case implementation and a dataset review was used. By keeping track of which cases an expert has actually viewed, traditional cornerstone case validation can be implemented. Any potential new rule is checked against cases this expert has previously reviewed to determine whether their list of classifications will change, providing exactly the same level of validation that traditional cornerstone case validation provides. However, the shift from case-centred to rule-centred thinking that this system can engender can cause problems here. It was found early in development that experts often considered many cases in the definition of a single rule, and approached the task from the perspective of defining a set of rules or classifications, rather than examining a single case and adding all the classifications that it should have. This can result in a

number of cases being half-reviewed, with some of their classifications added and some not, which would then result in a large number of cornerstone cases appearing whenever the expert eventually defined a rule that should have been defined for previous cases. To resolve this problem, cases were only recorded as having been reviewed if the expert explicitly stated that all the classifications for a case were present and complete, with the alternative option being to go back to the dataset and move on to another case. It is unclear from previous published studies whether any similar effect of half-reviewed cases is present in other MCRDR systems, but this might make for an interesting study and be a potential area for improvement. When viewing the dataset, cases that had been accepted as complete were highlighted in green to give a visual indication to the expert.

When these cornerstone cases were presented to the expert, they were given the option to view the case to verify what the classification for the case should be. Once they had determined this, the expert could choose to: accept that this case should have the new rule's classification; flag the case as having incorrect classifications, to be corrected later; or to modify the rule they are currently creating so that it does not apply to the cornerstone case. If they choose to modify the rule, the list of cornerstone cases that would be affected is updated and presented to the user as they add or remove new conditions. A rule cannot be added to the knowledge base until all of the relevant cornerstone cases have been either accepted, flagged as incorrect, or the rule changed so as not to create conflicts. When working with multiple experts on the same knowledge base, all relevant cornerstone cases were displayed whether they were marked as complete by the current expert or by another expert, as will be discussed in more detail shortly.

Additional Cornerstone Case Trials

Incorporating validation based on cases the expert has not yet seen is a more complex matter. While traditional cornerstone validation is helpful and effective, the unseen dataset should be a potential source of much stronger validation. As mentioned, whenever a rule is added to a knowledge base all cases in the dataset are checked and all cases matching the rule have that classification recorded. All case classifications are therefore recorded as the knowledge base is updated, meaning that the classifications for every case are consistent with the latest knowledge that

has been entered. It is also simple to identify all the cases in the dataset that will have a classification changed based on the new rule: all these cases can be considered to be de facto cornerstone cases.

This approach was implemented and trialled during development, with an iterative prototyping process and multiple experts. By treating unseen rule-matching cases as cornerstone cases, it was found that the validation became much less effective, as experts would become frustrated by the number of cases that would be presented to them. When over 10 cases were frequently presented in the validation phase (with occasionally numbers in the hundreds), some experts would ignore the validation process entirely rather than work through each case, even when the validation was in fact pointing out a relatively minor error. Even with a moderate number of cases some experts would become frustrated by having to look through them, especially when they were looking through cases which they had not previously seen.

To address this issue, the approach was modified by differentiating between ~~"t~~rue" cornerstone cases and unseen cases. However, this resulted in no improvement when large numbers of cases were presented, with experts likely to either ignore the unseen cases or still ignore the validation entirely. Common responses to the frustration of a large list of cases were to either assume that they had made a completely erroneous rule, and cancel the rule creation and begin again; that they had made some sort of mistake in past rule creation, and just click through accepting the new classification for all the cases without reviewing them; or that the system had made a mistake, and giving up on the process. The opinion was also occasionally voiced that reviewing them was unnecessary as they were confident in their rule as it was. The exact reaction seemed largely tied to the expert's confidence in their own abilities, but regardless of which of these responses was chosen, the result was negative and impacted the expert's interaction with the knowledge acquisition process. For these reasons the list of cornerstones displayed was restricted to only those cases that the expert had already reviewed and accepted as complete, with unseen cases forming additional validation through other means.

This implementation of the cornerstone system allows experts to receive validation feedback with a much larger dataset and define more complex rules than was possible with the EMCRDR system. The additional validation allowed by the use of

a dataset is implemented in a similar, though distinct manner, as described below in the Statistics section.

Statistics

The addition of a dataset provides potential for strong validation of rules, as there is more evidence to either support or refute the claims that the expert is making. However, presenting the evidence to the expert to be reviewed is problematic: the more evidence there is, the better the validation will likely be, but the more difficult and time consuming the task becomes. This is unfortunately a significant concern, as was seen in the execution of this study: when the task becomes too time consuming there is a tendency to skip details and ignore problems, leaving the validation ineffective. Because it is impractical to force an expert to review all cases that may be of interest, and likewise it is impractical to show these cases in any lengthy format, summaries and statistical information on groups of cases were used to provide the additional validation. To avoid the problems with expert frustration it was necessary to make the statistics an optional feature of the process: dataset statistics were always presented, but the experts were never forced to review them to make their rules, unlike explicit cornerstone case validation.

When defining a rule, the statistics presented related primarily to the set of cases that were covered by the current rule conditions, taking into account the position of the rule in the knowledge base tree structure. Three sections of statistics were shown: classification coverage, describing the distribution of classifications for cases covered by the rule; attribute statistics, describing the maximums, minimums, means and standard deviations for the cases covered by the current rule; and the best correlated attributes (up to 10). Also included were options to view the cases for each classification, either those within the set defined by this rule, or all cases with that classification in the whole dataset. This was to allow experts to be able to easily identify if there were errors with their rule. The presence of any cases with a classification that was mutually exclusive (or unlikely to be concurrent) with the conclusion of the rule currently being defined, should indicate to an expert that there is a deficiency in the rules somewhere. Likewise an expert should be able to identify unexpected proportions of other classifications, and examine the cases to update the rules accordingly. This simple mechanism allows an extra layer of

information about the interaction between classes and their attributes, potentially providing additional validation during the acquisition of knowledge. To enhance this feature, comparisons were made between the frequencies of each classification in the current case set and the overall dataset, and statistically significant differences were highlighted (For a more complete description of the statistics displayed, and the methods used to determine significance, refer to section 4.2.2 of this study). The statistics give a simple visual indication of which classifications are somewhat related to the classification currently being defined. This provides a small measure of extra validation, provided the expert has an understanding and expectation of which classifications should be related in the set of cases being used.

Similarly, the attribute statistics are intended to provide extra validation, providing the expert has an expectation of ranges and averages for certain attributes. However, this is a large list to look through, and is less likely to assist in validation for most experts as they will not have detailed expectations for the values of each attribute. As the attribute statistics also take some computation time, these statistics are not displayed by default in the system. The correlated attributes section summarises the ten attributes with the most significant difference comparing the values in the current rule's set of cases with the entire dataset. Like the main attributes statistics, these can provide assistance to validation but are not shown by default. As the attribute statistics are primarily used in the knowledge discovery sections of this study, they are described in more detail in Chapter 4.

Working with Multiple Experts

The last change from the standard MCRDR process was that knowledge was acquired from multiple experts, using different approaches. It was decided to employ multiple experts to help ameliorate the discrepancies and disagreements in opinion that can occur in the domain, and to make best use of the experts' time, as there were a number of experts available for only brief periods. Two methods were used to allow knowledge acquisition from multiple experts: having experts work collaboratively on a single knowledge base, in a similar manner to that taken by Richards and Vazey (Richards, 2009; Vazey & Richards, 2006); and having experts develop their own knowledge bases, which were consolidated afterwards, in a similar approach to that taken by Beydoun and Hoffmann (Beydoun, et al., 2005).

The former option had the advantage of requiring less work to develop a homogenous and complete knowledge base; the latter the advantage of keeping expert's opinions complete and distinct, and to make knowledge comparisons easier and more rewarding. Comparisons of expert knowledge are discussed in more detail in section 5.3.1.

Multiple experts with a single knowledge base

As described in section 2.2.2, this is the standard approach that has been used for acquiring knowledge from multiple experts (Compton & Edwards, 1994). However, problems can arise in this approach when the experts involved disagree on a rule or classification. Two types of problems were considered: situations where a less knowledgeable expert made a mistake that a more knowledgeable expert could correct; and situations where equally knowledgeable experts disagreed on what the correct rule or rules should be. The differentiation between these two, from the perspective of finding the correct knowledge, can be quite minor: often it can be reduced to whether one expert feels comfortable or feels pressured to accede to another's authority. It is often impossible to verify the truth of knowledge, and no matter how well learned they may be any expert can make a mistake. The correct knowledge may even be very different from either expert's opinion. In either case, it would therefore be best for the method to check the validity of each expert's opinion as much as is possible; although from a purely methodological perspective, the correct knowledge will ultimately be whatever the experts decide is correct.

Identifying Novice Errors

The traditional error-discovery strategy, whereby it is assumed that any error will eventually be discovered and corrected, is still in use as it is an automatic function of MCRDR knowledge acquisition. It is to an extent an unavoidable aspect of RDR: errors when defining a rule are likely to be noticed only once that error causes the incorrect classification of a new case. This process is improved by having multiple experts work with the knowledge base, particularly if both experts review the same cases, but also because a new expert is more likely to notice another expert's mistake. For the situation under consideration, with one expert correcting a less knowledgeable expert's mistake, having the knowledgeable expert review all the cases that the other reviewed would give the best chance of discovering errors. This

is less wasteful than it first appears as for the most part, if their knowledge is at all similar, the expert will only need to confirm that the classifications are correct rather than go into any detail of defining rules. It is however time-consuming, and can become frustrating and repetitive. Given restricted availability of experts in this study, this approach was not taken in order to maximise the knowledge acquired. If enough cases are evaluated then the traditional error-discovery approach should be sufficient.

Resolving Novice Errors

Of the two scenarios discussed, the key factor differentiating them is whether one expert is happy to accept the other's opinion as more likely to be correct. While this may be attributed to politics, prejudices, and perhaps even sheer stubbornness, much of this can avoided by simply ensuring that each expert knows who is disagreeing with them. Traditionally, experts will be working on the system within the same physical location and will likely have occasion for discussion. Whenever an expert notices another's mistake, they can correct it in the system, and explain the mistake in person; or, more likely, an expert will be aware of their standing in the work and will know whether to accept other's corrections or to make corrections themselves.

However, with online technology becoming ever more prevalent, and with this system being developed online, experts could work on the same knowledge base from almost any geographical location, thus invalidating the assumption that the experts know each other and can discuss matters in person. It is expected that the impact of this would be unpredictable. Depending on personality and confidence, when faced with an unknown person disagreeing with their statements an expert may be inclined to accept or reject the change without due consideration of why the change was made or how likely the change is to be correct. It is unknown whether Vazey and Richards made any such findings with their collaborative developments. At the very least, under this approach there is no mechanism in place to correct the mistake in the minds of the experts.

As with the collaborative approaches of Richards and Vazey, in this study it was always noted which expert defined each rule, and the creator was displayed when the rule was used to reach a classification or when the rule was examined. As no

deletions or edits of rules were allowed however, following the principle that knowledge is only added to the knowledge base, the ―change history" feature was unnecessary. Provisions were also made to ensure that when experts were working on a single knowledge base they were each made aware of the identity of the other participating experts, and provided contact information in order to make informed decisions in this regard; although in actual knowledge acquisition it happens that the collaborating experts were already acquainted.

Identifying Conflicts

In resolving conflicts, the first step is to identify that one has occurred. Conflict identification is mostly handled in one direction by the MCRDR rule validation, in that it will automatically notify the user when their new rule changes a previously accepted case. However, if the current expert disagrees with what a previous expert has said, that previous expert will need to be notified of this before the conflict can be resolved.

As with the identification of errors, an expert would normally be notified that another has disagreed with them either by eventually encountering a case where the same conditions apply, or by discussing it in person. However, in this study an extra measure was taken to display disagreements to the experts. Any case which an expert had previously accepted (and had therefore been added to the list of potential cornerstone cases), whose classifications had been modified by another expert, would be highlighted in red (as opposed to the green typically used to display a case that had been accepted) and displayed to the expert when they viewed the dataset. An option was also provided to select all the currently marked cases. This list of conflicting cases was also kept available to the administrator of the system, to additionally monitor conflicts.

Resolving Conflicts

As with the error resolution, the approach of assuming that conflicts will be resolved in person is flawed: experts may well work in different locations and have no knowledge of other contributors. As with Richards and Vazey's collaborative approaches, mechanisms for resolving conflicts in this study mostly involved creating a dialogue between the experts. It was ensured, with participant approval, that the experts were able to contact each other online about disagreements. In

addition, the current conflicts were monitored by the administrator of the system, and where no other progress was being made the administrator would attempt to reconcile them via discussions with as many participating experts as possible.

The major flaw with this approach was that it assumed a lengthy commitment of continued use of the system by the experts involved. As mentioned previously this was not possible in this study, with some experts not even able to work on the system concurrently. This significantly reduced the likelihood of any dialogue through the system, and reduced the likelihood of experts identifying conflicts or errors, limiting the efficacy of the collaborative knowledge acquisition in this study.

Knowledge Base Consolidation

The alternative approach is to allow each expert to work on their own knowledge base and attempt to reconcile them afterwards. Given restricted time with experts, a similar integration approach to Beydoun and Hoffmann's (Beydoun, et al., 2005) was also taken, although with less focus on automatic integration and more on evidence-based conflict quantification and resolution.

Reconciling Knowledge Bases

The process for reconciling knowledge bases was based on comparing the results for each knowledge base over the dataset, keeping any comparison grounded with evidence, and resolving the conflicts with discussion between the experts. In order to allow these comparisons, similar classifications were grouped into equivalencies where necessary, through a simple interface of selecting the equivalent classifications and marking them as a group for comparison purposes. These groupings were decided through consultation with the experts, in order to manage the different terminology and levels of detail that different experts might use. More detail and examples are given on exactly how these comparisons were performed in Chapter 5.

The comparisons of results over data identified which classifications had different definitions between experts, including the magnitude of each conflict. Once these were identified, the administrator of the system contacted the experts involved, initiating discussions to resolve the conflicts. When presenting the conflict to the experts, the cases which had different results provided an easy way to describe the exact context in which the conflicts arise, regardless of how complex the rules may

be that lead to them; and as per the RDR philosophy, the provision of context is an important consideration in consulting experts and receiving useful responses.

Summary of Modifications

The four modifications made to the process are repeated here for clarity:

- Presenting a dataset, not a case
- Defining potential cornerstone cases based on expert acceptance
- Dataset statistics for additional validation
- Working with multiple experts: collaboration and consolidation

Each modification is somewhat tangential to the central knowledge acquisition process, as the interpretation of individual cases, and the definition of rules and classifications are all unchanged. However, each modification can still bear an influence on the efficacy of the knowledge acquisition, as is shown by the results and discussed in the following section.

3.3 Results and Discussion

One of the most interesting findings related to the comparison and consolidation of the two knowledge bases. The comparison method and results are covered in detail in Chapter 5, as it is more pertinent to that discussion. Initially this section will discuss the resolution of the conflicts identified by that comparison, including such comparison details as are necessary for context. Following this is a detailed presentation of the development and performance of the amalgamated knowledge base as an expert system.

3.3.1 Knowledge Base Consolidation

In consolidating the knowledge bases, there were few instances where the experts clearly disagreed with each other; most of the differences could be accounted for by small inconsistencies between the knowledge bases. For example, one knowledge base might use *greater or equal to* as a border between classifications, whereas the other uses *greater than*. Differences such as this were also quite common within each knowledge base, for example the rule for one gradation of classification might use *less than 40* while the rule for the next level uses *greater than 40*, excluding any case which falls exactly on 40. The resolution to these border definition problems

were typically fairly arbitrary, as the experts tended to consider it unimportant which resolution method was used, as long as it was consistent. Each occurrence was of little consequence individually, but accounted for 1256 cases (42.4% of the dataset) reaching different classifications overall. Once resolved by the administrator these differences had little further effect.

Another source of disparity resulted from the manner in which the knowledge bases were built: one expert might add a rule as an exception, where another expert adds that same rule at the root level. If the first expert does not encounter any rules which should match the exception, but do not match the rule it is an exception to, then the mistake will not be noticed.

There were five identified instances of distinctly different definitions for classifications that did not have obvious solutions. To resolve these, the administrator initiated an email conversation including each of the two experts involved (the primary experts). In one instance both considered the other's opinion too extreme, so a compromise was found that both experts were satisfied with. On another occasion one expert declared he had no objection to removing the differing condition from his rule. For one other conflict, the primary expert explained his rule difference to the other expert, who happily accepted the change once he understood.

The final two differences were not able to be resolved by discussion of rule conditions alone. For one of these, one expert had included in the collaborative knowledge base an alternative rule for reaching a given classification, which had no counterpart in the independent knowledge base. Although confident that the rule was not an accurate definition, the other expert could see some logic behind it and was open to the possibility that it may be useable. In response, the expert who had added the rule suggested looking at how often the rule misclassified cases compared with the other, more widely accepted rule. Using the statistical tools implemented for the knowledge discovery section of this study (discussed in Chapter 4), it was found to give 158 false positives and 9 false negatives, with 7 cases matching both rules, out of the 1390 cases in the TAHS dataset. This was deemed far outside the expected parameters for intended coverage of the rule; hence, although the alternative rule may have correctly classified some cases, the inaccuracies were deemed too great and the rule removed.

The second conflict arose from a condition added to a rule by the expert in the independently developed knowledge base. The rule existed in both knowledge bases, in the same form except for that condition. The condition was added to cover a small contingency of cases that the expert considered a possibility. On learning that no other expert had included such a condition however, the expert expressed some uncertainty and a feeling that the other expert may in fact be better educated in this instance. To resolve the problem, he requested data on how many cases were affected. Given that only 15 of 947, or 1.5% of cases with the classification also matched on that condition, the expert decided that any potential benefit did not outweigh his uncertainty and decided to remove the condition.

Resolving these conflicts led to the resolution of 613 cases (20.7% of the dataset) that had previously reached different classifications between the knowledge bases.

The equated classifications provided a simple way of comparing the results of the knowledge bases, to identify problematic differences. In consolidating the equivalent classifications into a common structure, one set of classifications (the *Obstruction* group) were problematic: each knowledge base used different versions of the classifications, both in terms of gradation and in compound classifications with another classification (*Reversibility*, or, *Positive response to bronchodilator*). The experts were consulted as to which of the gradations of severity should be used, and which definitions to keep. It was considered relatively unimportant, the end result being much the same in terms of providing a sufficient interpretation; and the version included in the initial documentation was kept, as that document had been circulated and confirmed by other experts. The *Reversibility* elements were separated into separate classifications, at the assurance of the experts that this was not problematic or any less correct.

Each of the other groups were relatively simple to consolidate, once common differences were resolved. In each case the more detailed versions were included for completeness.

The numbers of conflicts presented here highlight that a standardised knowledge base and an expert system can be very beneficial to the domain: even between 3 experts, the two major contributors of which are high-ranked specialists in the field,

1869 out of 2963 cases (63.1%) received different classifications. Of these, 613 (32.8% of the conflicts) were major and needed intervention to resolve.

The very minimal input from the secondary expert in the collaborative knowledge base unfortunately precluded any comparison between collaborative approaches and post-acquisition compilation of knowledge bases.

3.3.2 The Expert System

The resultant amalgamated knowledge base forms a functional expert system for the lung function domain, which is of benefit to the domain in assisting experts make consistent and complete interpretations. It also provides a knowledge base capable of providing guidance for complex data analysis, as will be discussed in the following chapter. This section will examine the details of how this expert system was developed and how it performs.

3.3.2.1 Accuracy

As has been noted previously (Bindoff, 2010), the accuracy of an expert system such as this cannot be directly measured without extra, pre-classified data: the system is always correct on every case it has already seen, and asking an expert to classify cases outside of the system in order to test it seems a waste of expertise and expert time. However a measure can be found, by considering the number of corrections which an expert needs to make as they are examining the system. This has been described by the formula from Bindoff's work presented below (Bindoff, 2010):

$$accuracy = \frac{\mathrm{Cf} - \mathrm{Crem} - \mathrm{Crep}}{\mathrm{Cf} + \mathrm{Ra}}$$

This equation provides a measure for how accurately the system classifies each case, as it is interpreted by the expert. *Cf* is the number of classifications initially found for the case, *Crem* is the number of classifications removed by the expert, *Crep* is the number of classifications replaced, and *Ra* is the number of rules added by the expert for the current case. Assuming that the expert completes each case before moving on to the next, as more cases are seen there should be a trend towards increased accuracy. This should plateau as the knowledge base approaches complete coverage of expert knowledge.

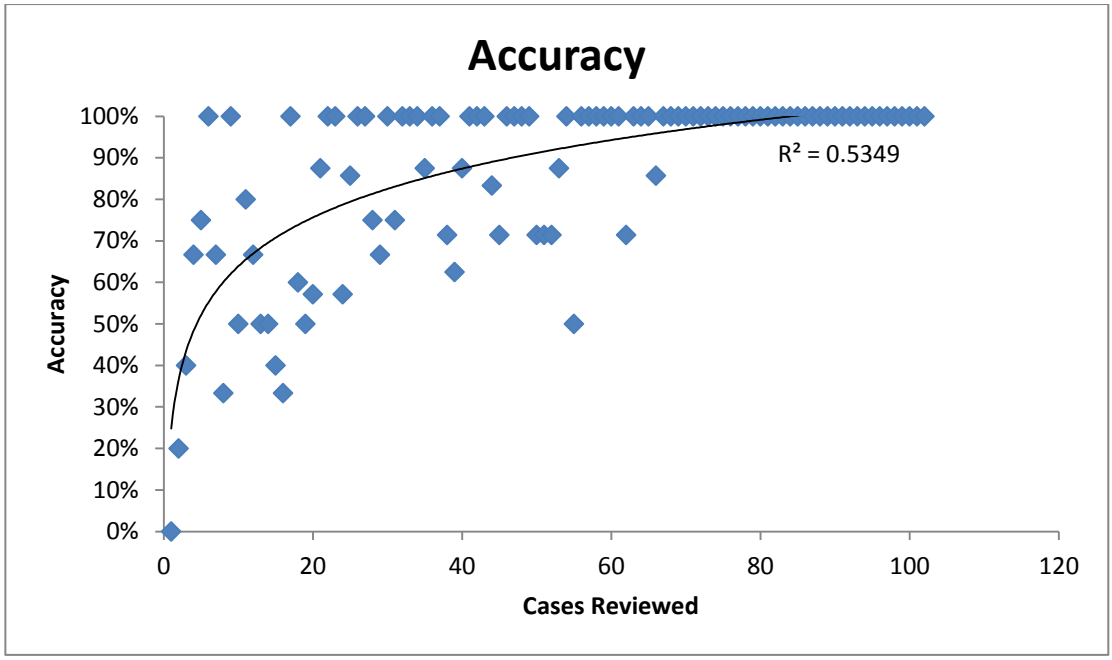


Figure 3-4: Accuracy of system as more cases are reviewed

Figure 3-4 shows the accuracy of the system over the 102 cases reviewed by the independent expert. The trend line, while certainly not a perfect predictor, gives an indication of the general pattern: a quite rapid rise in accuracy which quickly slows as more cases are seen. The trend line suggests that perhaps the plateau has not been convincingly reached, as there is still a noticeable positive slope at the end of the line. However, as the knowledge base achieves 100% accuracy for the last 36 cases (over one third of the cases seen overall), it seems reasonable to assume that there is little if any improvement left to be made. Certainly, after reviewing 36 cases without having to make any changes, the expert was satisfied that the system was complete and had little patience to continue.

3.3.2.2 Rule Creation

Rules per Case

The number of rules created for each case examined gives a good estimate of the rate at which the system is acquiring knowledge. Figure 3-5 shows the number of rules added per case examined, in the order that the expert examined them.

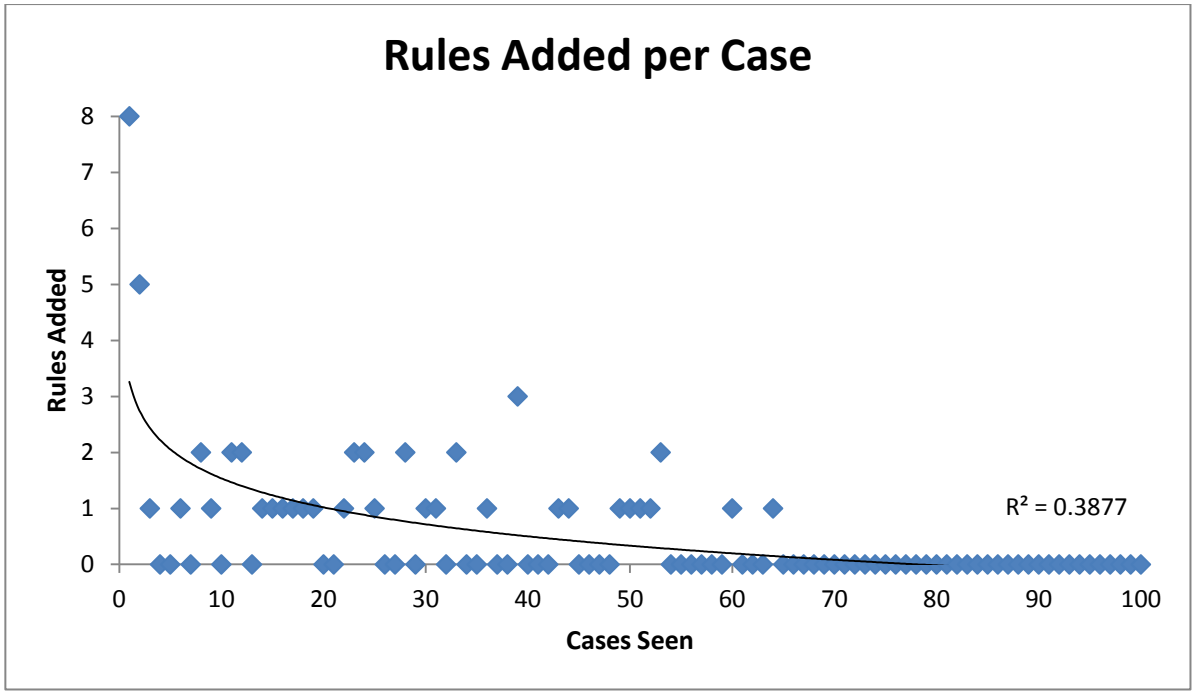


Figure 3-5: Rules added per case

The trend line, while not particularly representative of the data, gives a general indication of the pattern: after an initial very high number of rules added per case, the numbers steadied, with a gradual slowing as more cases were seen. As the number of rules added only once went above 2 after the initial few cases, and only once went above 1 after the first 50 cases, the rate of rule addition appears to be consistent within this downward trend.

Time per Case

The time taken to complete each case can be an indicator of many different aspects of the knowledge acquisition process, such as how detailed the expert is being in interpreting the data, how complex the task is, which cases prove difficult, how they adapt to the system interface, and the variability in complexity in cases. Figure 3-6 shows the time taken for each case reviewed. As would be expected, both from the expert increasing in familiarity with the system and with the system improving in accuracy, there is a trend of continued improvement in speed as more cases are seen. The average of 3 minutes and 29 seconds well represents the data, as despite the obvious upper outliers, the majority of cases are in the lower section, with 65% of cases falling below 3 minutes. The standard deviation of 3 minutes and 33 seconds

does demonstrate just how variable the times are however, highlighting that interpreting cases is not necessarily a simple, consistent process.

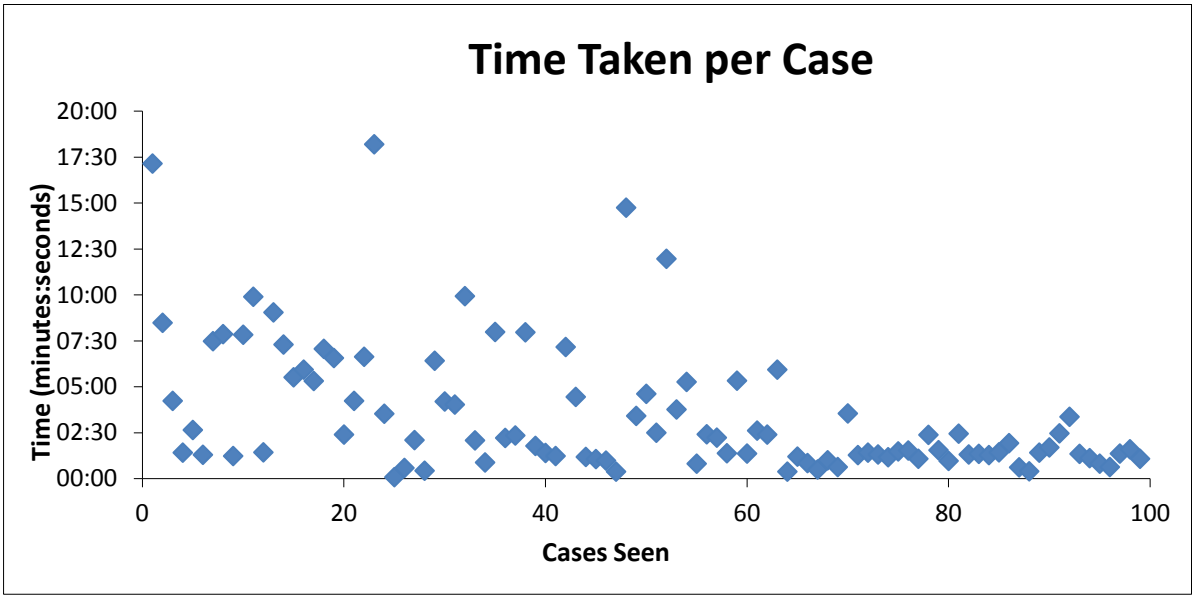


Figure 3-6: Time taken per case

Time per Rule

The time taken to create each rule gives an indication of the expert's familiarity and ease with the system, the complexity of the knowledge being defined, and the number of problems the expert faced in defining a valid rule. The principal limitation with this measure is that it is difficult to distinguish which of these factors are having the most influence on the data.

Figure 3-7 shows the time taken per rule created, in sequence as they were created by the expert. The moving average demonstrates the variability inherent in the rule definition, which is to be expected: some rules are more complex to define than others; and the process of finding an explicit definition of tacit knowledge can be a variable, incremental process, depending on how long it takes to find a definition that fits. The average time taken per rule is 2 minutes and 29 seconds, a fairly typical number for systems of this kind, but with a standard deviation of 2 minutes and 6 seconds which further highlights the variability. The data is slightly suggestive of an upwards trend, but with little conviction. The relatively low average time for initial rules is somewhat expected, as the initial rules added to a knowledge base are typically quite general, classifying broad segments of the

domain, and are the rules which the expert uses most routinely. Hence these rules are generally quick to define. However, this can be balanced by the inevitable period of acclimatisation to the system interface: learning how to view the cases, how to define the rules, and learning the particular details of the data model. This acclimatisation period likely accounts for the large initial variability in the first few rules: the logs indicate that 70% (approximately 1 minute and 40 seconds in each) of the time spent defining the second and fourth rules was involved searching through the attribute and classification lists to try and find the appropriate entries; a task which took considerably, and increasingly, less time for future rules.

That the time required to define rules does not decrease might be explained by increasingly complex rules being entered. This is supported by Figure 3-8, which shows the number of conditions per rule increasing over time. Both time per rule and conditions per rule follow similar patterns, suggestive that the complexity of the rule is a strong influencing factor on the time required. The addition of two zero-condition rules later in acquisition also supports that the rules are becoming increasingly complex, as these were required to remove incorrect rules: the presence of which is a good indication of the complexity of the task.

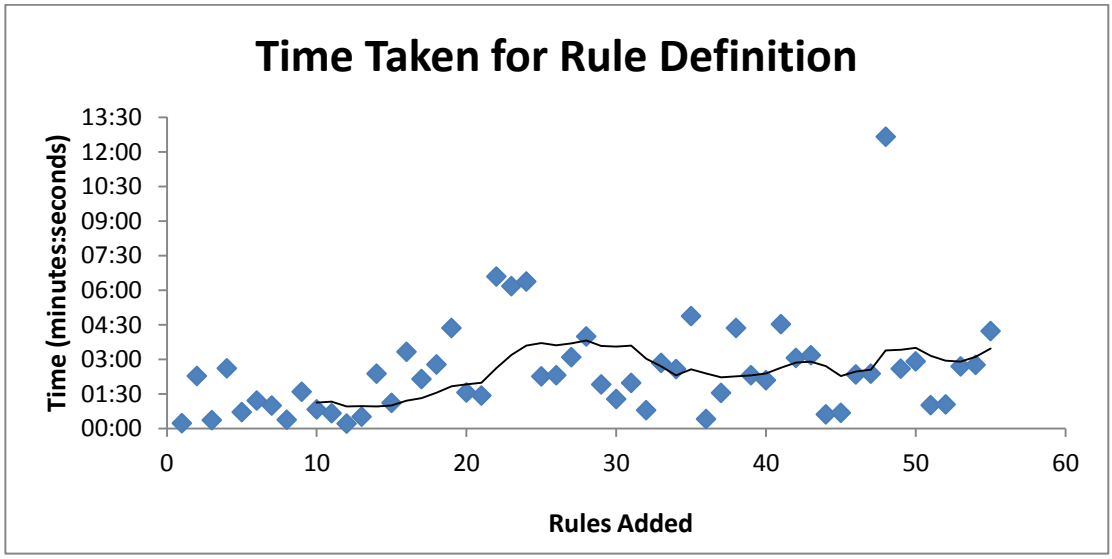


Figure 3-7: Time taken to define rules, with a 10-based moving average

The maximum time taken for a rule was just over twelve minutes, and occurred quite late in the knowledge acquisition process: an examination of the logs indicate that four minutes of this time was spent reviewing the two cornerstone cases that

were identified, with the final result of accepting the new classification for both of them; a further six minutes were spent deliberating over the values to be used in the conditions; whereas the classification was decided upon without change within seconds. This would suggest a classic example of defining a rule from tacit knowledge (Richards & Busch, 2003): the expert knew exactly what the classification should be but took some time and effort expressing why this should be the case, and exploring how to precisely articulate the differences between the new classification and previous classifications.

The second longest rule to define, at six minutes, found no cornerstone case to examine in rule validation with most of the time taken searching for the desired classification and attributes in the interface. The third and fourth longest rules, again at six minutes, had similar difficulties with identifying the desired classification, with most of the time taken by examining cornerstone cases and modifying conditions to validate the rule.

Conditions per Rule

The number of conditions used per rule gives a reasonable indication of the complexity of knowledge being added, with more conditions generally indicative of a more complex rule. The numbers displayed may be misleading in this domain, due to the prevalence of attributes calculated from two or more other attributes.

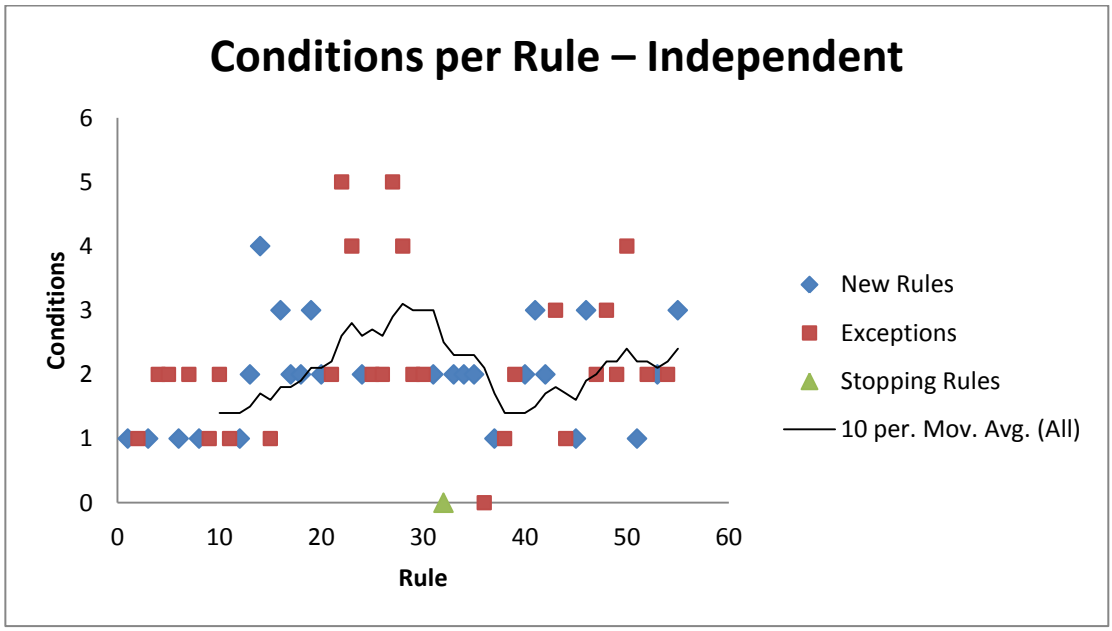


Figure 3-8: Conditions per rule for the independent knowledge base, with new root-level rules, exception rules and stopping rules identified

Figure 3-8 above shows the frequency of each number of conditions. The graph shows that the majority of rules (40 out of 55) had only one or two conditions, indicating that the rules are typically fairly general; although the tendency to use ratios between two attributes as a single attribute contribute to this slightly. Of the 113 conditions defined, 37 used compound attributes, and nearly every other used a *percent of predicted* attribute, formed by comparing a measured value to a predicted value.

The average number of conditions used per rule by the independent expert is 2.05. As mentioned previously, the data could be said to show an upwards trend, particularly discounting the zero-condition rules, suggestive that the experts were attempting to define more specific and complex rules. There is a significant drop in the average between the 30 and 40 rule marks, as two rules are added with no conditions (correcting previous errors by ensuring that the old rule will never again be able to fire).

Interestingly, there appears to be a very even spread of exceptions and new root-level rules. The exceptions in this study cannot all be said to be correcting errors, as it was suggested to the experts that they define general classifications first and use exceptions to refine them into more specific sub-classifications. Figure 3-8 shows a reasonably even spread of new rules and exception rules over time, with perhaps a slight increase in exceptions towards the end of the acquisition, which would be expected as most common rules have been added and errors in previous definitions are encountered. The mean number of conditions per new root-level rule was 3.8, whereas the mean for exceptions was 2.2. Exceptions are expected to have a lower number of conditions as they are often small refinements of existing rules, and this seems to be reflected in these averages.

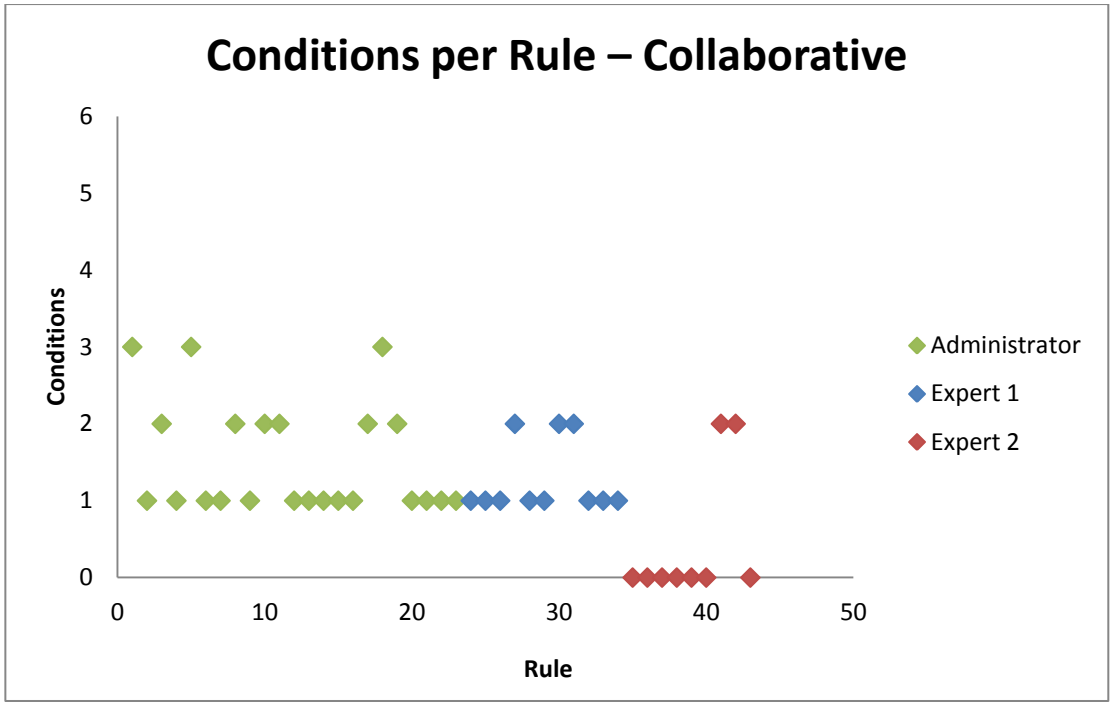


Figure 3-9: Conditions per rule for the collaborative knowledge base

Figure 3-9 shows the conditions added for each rule in the collaborative knowledge base. The rules added by the administrator are those defined in the initial documentation developed by the expert that worked on the independent knowledge base, and were agreed upon by a group of experts. It can be seen that these rules generally conform to the complexity of the initial rules defined in the independent knowledge base, with mostly single condition rules and some dual condition rules, with very few more complex than that. This is expected, as the rules listed in the document consist of a summary of the general knowledge in the domain. There are however more of these simple initial rules defined here than in the independent knowledge base. This is likely because the administrator entered all of these rules in the order that the rules were presented, rather than by waiting for an exemplar case to be presented, which would cause these initial simple rules to be defined over time in the independent knowledge base. Of note however is that the subsequent rules in this knowledge base stay at a similar level of complexity throughout, with no especially complex rules being defined. It is interesting that the other two experts also did not define any more complex rules. This may have been an attempt to conform to the complexity of knowledge of the rules already defined; it may also be a factor of the collaborative nature of the knowledge base, with experts unwilling to

define complex rules that would be subject to scrutiny by other, potentially more knowledgeable, experts.

Coverage of Rules

As this study uses a dataset of cases, which are classified as each rule is added, it is possible to determine the rule coverage as each rule is defined: in other words, the number of cases classified by each rule when it is added. Assuming that the dataset is representative of the domain and does not contain an unreasonable distribution of case types, this can provide a measure of the specificity or generality of each rule. It is expected that the expert will begin by defining fairly general rules, and as these are established the rules will become more complex, as the expert attempts to deal with more detailed classifications and to resolve inconsistencies with previously defined rules. Figure 3-10 shows the number of cases classified by each rule, in the order that the rules were added to the system.

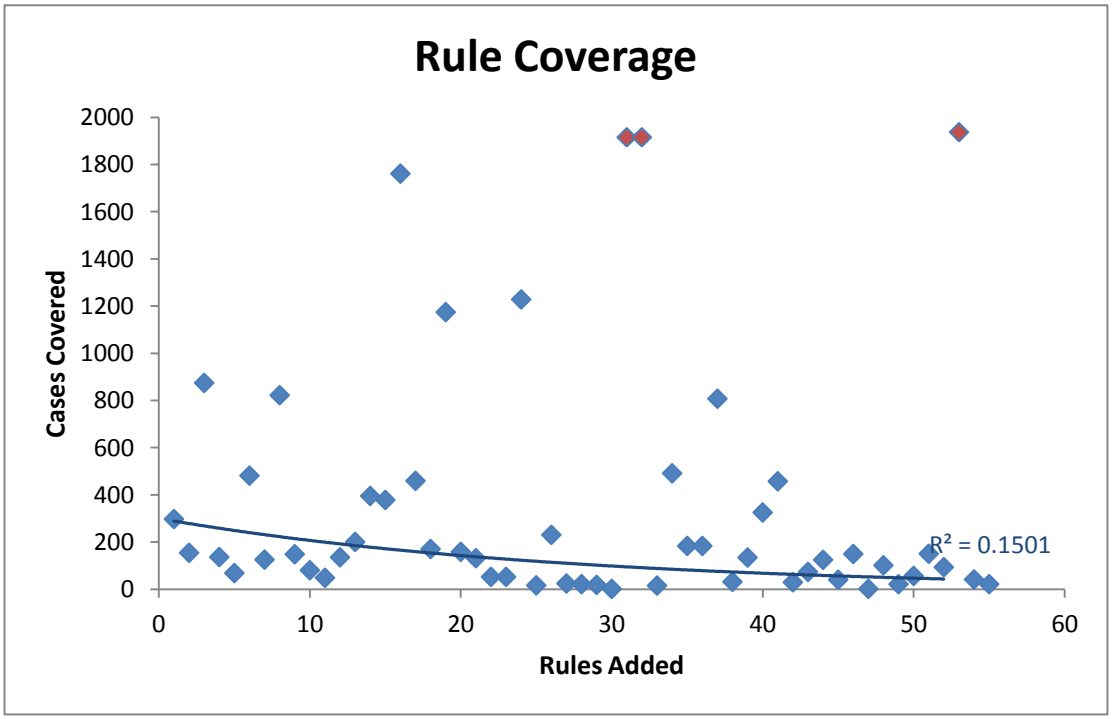


Figure 3-10: Number of cases covered by each rule, in the independent knowledge base, with identified outliers indicated in red

The graph does perhaps indicate a very tenuous slight downward trend in the number of cases covered over time. It does appear obvious that a majority of the latter 50% of the rules have a smaller coverage, with 21% of the first half of the

rules covering less than 100 cases compared to 48% in the second half. The data is also skewed somewhat by 3 outliers in the second half, marked in red on the graph. The first two of these represent a mistakenly added rule whose coverage was far too broad, which was then immediately removed by a stopping rule again covering all the same instances. The third outlier is a very general classification the expert decided to add towards the end of the knowledge acquisition process, which covered many of the cases already seen – meaning that it would have been added within the first few cases examined had the expert always intended to include that classification. Removing these outliers from consideration gives a better visual appreciation of the trend of changing rule coverage as the knowledge develops.

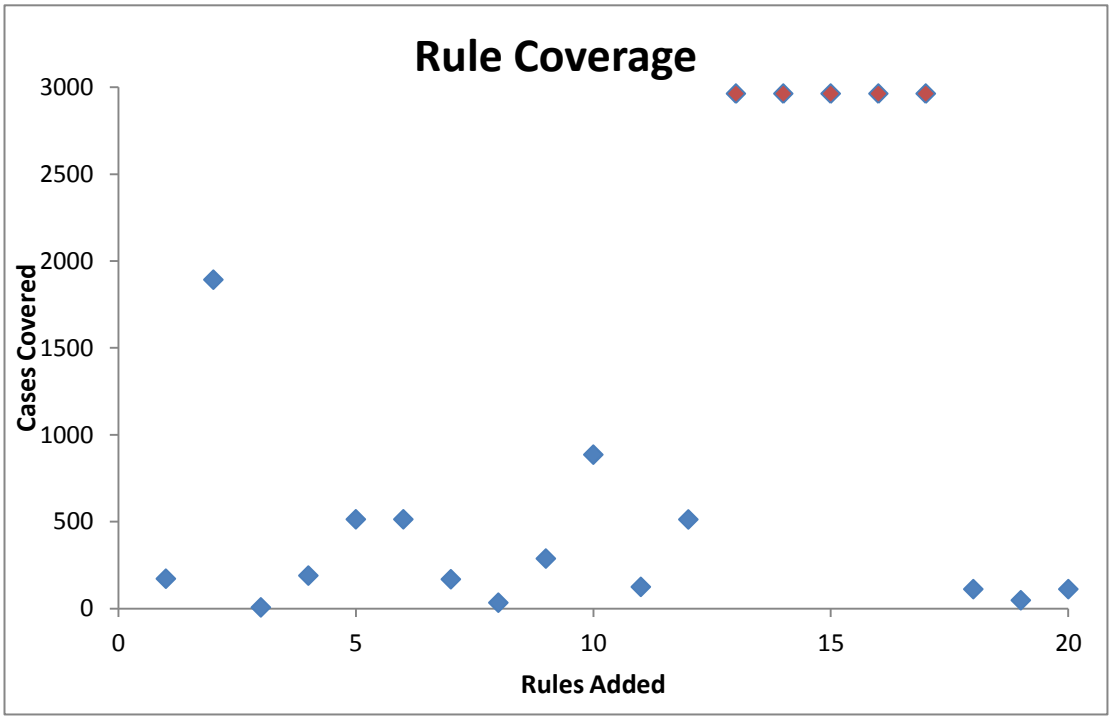


Figure 3-11: Number of cases covered by each rule, added the by the second and third experts, with identified outliers indicated in red

Figure 3-11 shows the case coverage for rules added by the second and third experts to the collaborative knowledge base. Again the red marked rules are excluded from the trend, as they constituted rules with no conditions that covered every case, entered by error. A similar downward trend in cases covered may be occurring, as the figure shows quite similar ranges of coverage compared to the other knowledge base; but the data is inconclusive. Similarly most rules cover 500 or fewer cases.

3.3.2.3 Classifications

Number of Classifications

The independent expert defined 29 distinct classifications in addition to the 21 classifications that were initially added under this expert's guidance. Of these 50, 43 were used in classifying cases; with 54 classification-reaching rules, this gives a ratio of 1.26 rules per classification, suggesting that the expert performed well in quantifying each classification into a single, general rule.

Classifications per Case

The number of classifications made per case can give an indication of the level of detail that the expert uses in describing each case. Figure 3-12 summarises the frequency of the number of classifications each case received, for both the independent and collaborative knowledge base (before consolidation). It shows again that the independent expert went to more detail than the collaborative knowledge base.

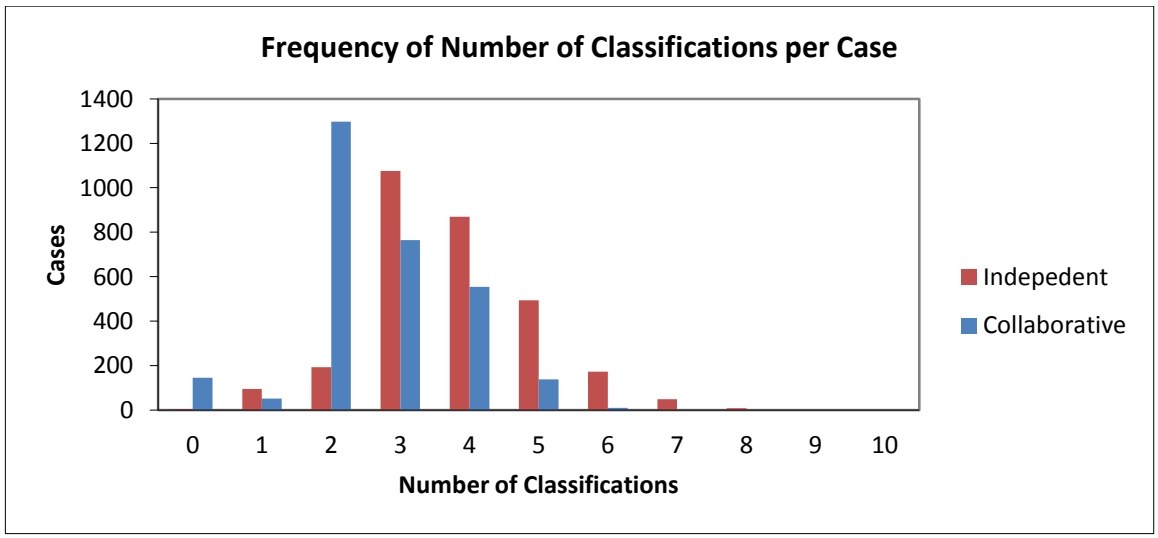


Figure 3-12: Numbers of cases having each quantity of classifications

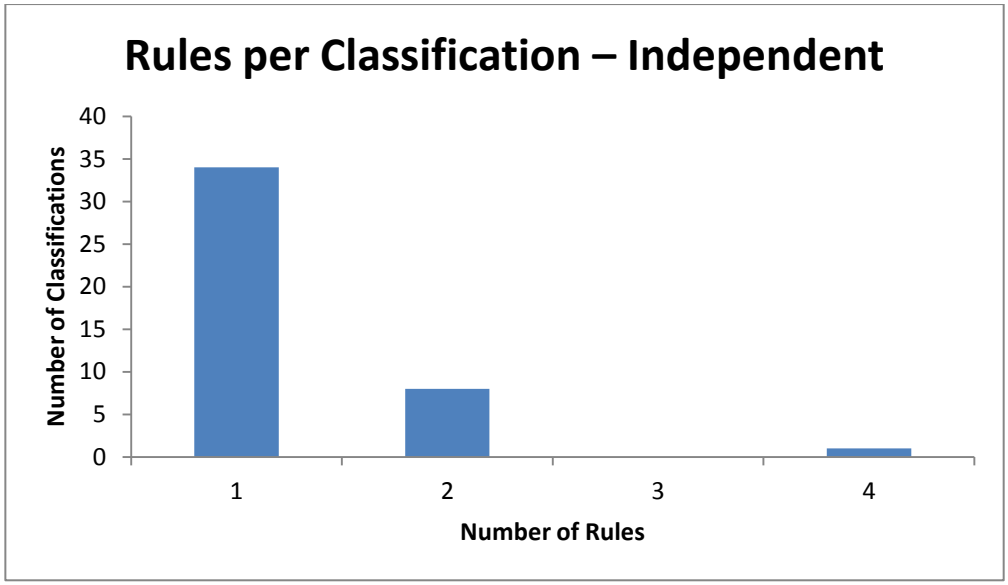


Figure 3-13: Frequency of number of rules for each classification

Rules per Classification

Figure 3-13 shows how many rules were added per classification, in the independent knowledge base. It clearly shows that most classifications had a single rule to derive them, with an average of 1.26 rules per classification. Of the 43 classifications used in this knowledge base, only 9 had more than one rule to derive them. It is worth looking at some of these in more detail to understand what these nine represent.

Normal Spirometry

This classification is the most unusual of the 43, and it is the only classification to have more than two rules, actually having four rules devoted to it. On examination of these rules however it is immediately apparent that the expert had difficulty in explaining this classification. The first rule seems straightforward, with seemingly reasonable conditions, except that it has an attached exception rule which also has the same classification *Normal Spirometry*; and is in fact the second rule. This will have no effect on the results of the system: any case which matches the first rule, and matches its child exception rule, will end up with the same classification being applied, and no measurable difference to the user. That the rule was added is simply due to the expert misunderstanding how the knowledge base structure works: having realised that the rule he just added was missing a condition, the expert

attempted to correct the mistake by immediately "changing" the existing classification and adding the further condition. As the 13th rule added it was still relatively early in the knowledge acquisition process. This error was corrected in the consolidated knowledge base, but still serves as an example of how the expert interacted with the knowledge acquisition process.

The third rule constitutes a refinement of the first rule, in that it supersedes the first rule entirely (at least when the first rule is considered as a combination of the first and second, as was intended by the expert). The conditions are identical to those in the first rule, except for the first condition which is modified to broaden the coverage of the rule, and to provide consistency of coverage with other rules that had been defined.

The fourth rule represents a different event. It is much more general, as it has two of the same conditions as the previous rule but is lacking the third, which would again cause it to supersede the previous rule and broaden the coverage of this classification. However, this can be shown to be an interpretation mistake by the expert: when the expert encountered a case that had been classified by this new rule (and not the previous rules), he noted that it was a misclassification and changed it, with the condition that had differentiated this fourth rule from the previous rules as the exceptional condition. The end result of this is that the rule covers exactly the same cases that the previous rule does.

While it could be coincidental that this classification happened to have more problems associated with it than any other, the reason may lie in the nature of the classification itself. For a medical expert whose job it is to identify the types and extent of the problems afflicting a patient, the definition of a single rule to define *normal* or *lacking any problems* can be expected to be a very difficult task. This should also be considered in the context of the MCRDR knowledge acquisition process: examine a case, describe your classifications, and then justify why you reached that conclusion. A likely thought process for such an expert will be to examine the case, and identifying the most likely problems. Investigating these further, if the expert finds that none of the problems are present, their classification will be *normal*. Their natural justification for this classification will be in terms of the problems they just ruled out: "The case is normal because, although it has some signs of problem X, it is lacking conditions A, B, and C." The rule they define may

then be specifically focused on certain problems not being apparent; or, even if they consider many possibilities, may easily miss some eventualities. Although the expert did establish a single rule to do this, it is unsurprising that they had difficulty and required some attempts.

Normal Lung Volumes

The second rule that was defined for the classification *Normal Lung Volumes* was added when the expert misclassified a case: as the case should not have actually had this classification, the second rule was overly general and was eventually stopped when the expert realised it was incorrect.

Fixed moderate obstruction

As with *Normal Spirometry*, this is an example of trying to define a rule for a classification that is not present. In this instance, *Fixed* refers to the case being not *Reversible*. As the rule for *reversibility* requires two separate conditions to be present, when the expert attempted to define the opposite rule they included the negative of both of those two conditions. However, the classification *Fixed* does not need both conditions to be present, as only either one of the *Reversible* conditions needs to be false, following de Morgan's law. This is an example of where allowing disjuncts in rule definition would have been useful. The second rule was added to cover some of the cases which the first rule missed, when a case was encountered which displayed one of the conditions but not the other; however evidently no case displaying the opposite combination was encountered, as the third corresponding rule was not added.

Mild airway obstruction

As with the second rule for *Normal Spirometry*, this is another example of mistakenly attempting to correct an existing rule by adding an exception rule with the same classification. The original rule was corrected in the consolidated knowledge base.

Mildly/Moderately/Severely impaired gas transfer

These three classifications required two rules due to the attributes used in the conditions. Cases can contain both a D_LCO value corrected for haemoglobin and an uncorrected value; or only an uncorrected value, or neither. The uncorrected value

should only be used where the corrected value is unavailable: necessitating two rules. This is another situation where disjuncts would have helped rule definition, provided conditions could also be grouped.

Severe obstruction

The second rule for this classification appears as an exception to an exception to the first rule for this classification: the expert defined an exception to the initial rule for *Severe obstruction*, but later realised that this exception was slightly too broad when he encountered a case that matched the exception rule but should have the original classification. He then added an exception to the exception, returning the classification for the case to the original *Severe obstruction*. This is a good example of the incremental nature of the knowledge acquisition, and how it will eventually discover details that the expert misses or is not explicitly aware of in their reasoning.

Moderately severe airflow obstruction

For this classification, it appears that the first rule defined was too specific. Eventually the expert encountered a case which should have had the classification but was not covered by the rule: and so another rule was added. These rules do not appear to be representing different knowledge, with the second rule encompassing the other, and so would not have been avoided by allowing disjuncts in rules.

The overall lack of multiple rules per classification would seem to indicate that the experts have a good understanding of the domain and are confident in their definitions for each classification. However, the occasions where extra rules are needed highlight the potential for expert mistakes, and the existence of the tacit knowledge the experts hold that are not included in these standard definitions.

3.3.2.4 Cornerstone Cases

Classifications per reviewed case

Figure 3-14 shows the number of classifications found for each reviewed and completed case. When the results of the independent knowledge base are considered in comparison to the classifications found for cases not yet seen, there is a striking disparity: the mean for reviewed and accepted cases is 5.1 classifications per case, whereas the mean over all cases is 3.8. This would suggest that the knowledge is far

from complete, as presumably there are still rules and classifications to be added for the unseen cases to bring their classification numbers level with those explicitly reviewed, despite the knowledge acquisition appearing to have plateaued in Figure 3-4.

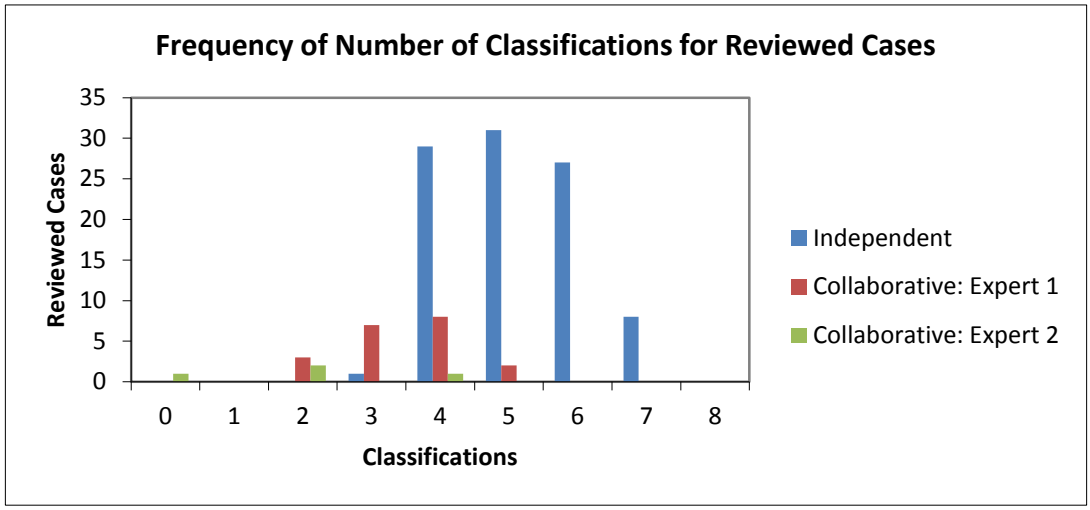


Figure 3-14: Number of classifications given to each reviewed case

A major contributing factor to this is the nature of the different datasets used. The TAHS data constitutes the testing of a broad population spread, regardless of their health; whereas the Austin Health and Royal Hobart data are taken from those people explicitly recommended for lung function testing because of their likelihood of health problems. The TAHS data therefore includes a strong bias towards healthy people, whereas the other datasets have a definite bias towards patients with problems. This is relevant to this analysis as in this knowledge base healthy people generally have less classifications than unhealthy: a common healthy set of classifications is {*Normal Lung Volumes*, *Normal Spirometry*, *No evidence of gas trapping or non-uniform ventilation*}, whereas an unhealthy patient will usually require more detail describing each problem. This tendency is described in Figure 3-15 and Figure 3-16, where it can be seen that the TAHS dataset has a substantially higher rate of classifications than the non-TAHS data, and the mean for the TAHS dataset of 3.3 classifications per case compared to the 4.2 for non-TAHS classifications.

However, even taking this into consideration, reviewed cases average one more classification per case than unseen cases. It was thought this may be accounted for

by the selection of the initial 20 cases for knowledge acquisition: a set which were chosen specifically to maximise coverage of classifications, and included some of the more complex cases for interpretation. However, those 20 cases show a mean of only 0.1 more classifications than the 76 other reviewed cases (5.2 to 5.1).

Considering Figure 3-14 it can be seen that of the reviewed cases in the independent knowledge base only 1 has less than 4 classifications. If it were to be assumed from this that cases should generally therefore have at least 4 classifications, this would show that almost half (1369 out of 2963, or 46%) of the cases in the dataset are missing at least one classification. While it would be rash to make such a conclusion, it is good evidence that the knowledge base, while seemingly complete after 96 cases reviewed, is still lacking in finer details for some cases.

This is not evident in the collaborative knowledge base, where the accepted cases have equivalent numbers of classifications to the unseen cases.

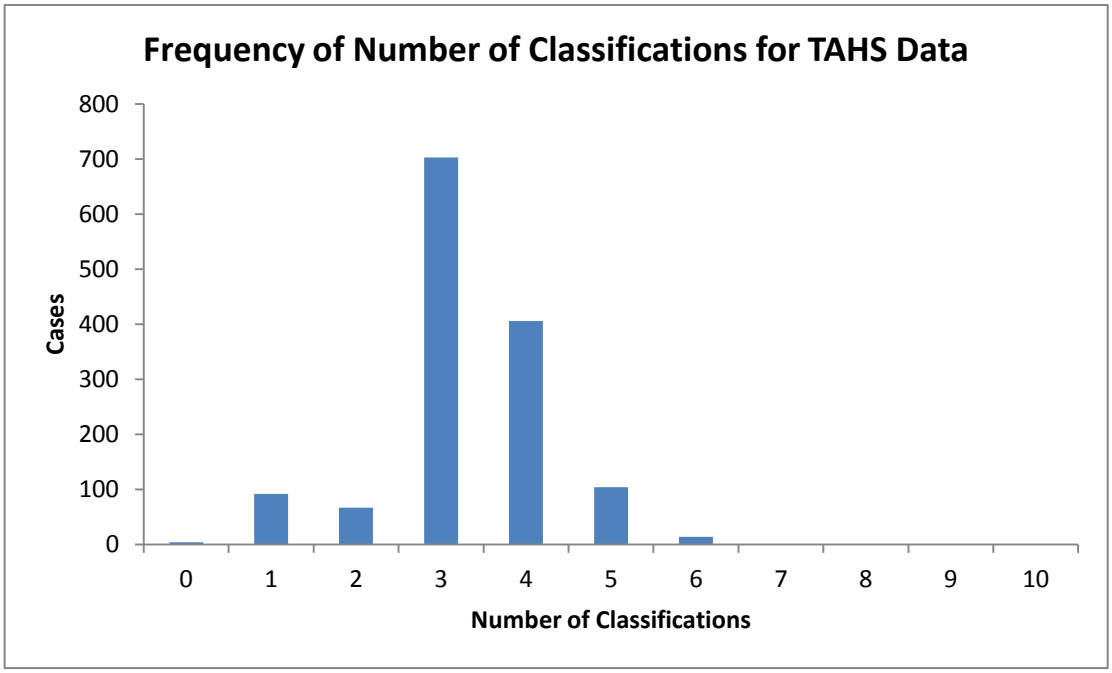


Figure 3-15: Frequency of number of classifications per case for the TAHS data

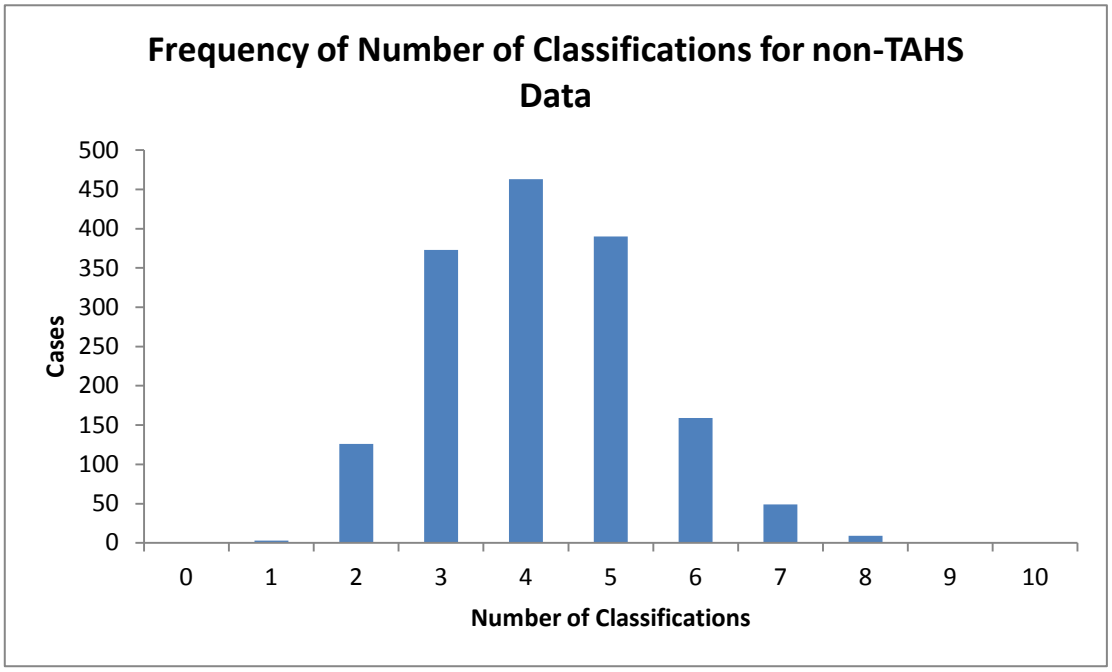


Figure 3-16: Frequency of number of classifications per case for non-TAHS data

3.3.3 Impact of Validation

The impact of the modifications to the validation process is difficult to measure directly, and is partially reliant on qualitative data from the experts on how useful it seemed.

3.3.3.1 Cornerstone Cases

Limiting potential cornerstones to only those cases which the expert has reviewed is sufficiently close to traditional MCRDR cornerstone cases identification that there is little resultant difference. The requirement that the expert choose to accept the case as complete, when they have the option to just move on, did have an effect however. Whenever the expert began classifying cases they would initially be hesitant to accept cases as finished, choosing to examine more cases in case they encountered something further they had missed, or had misunderstood some action in some way. The result of this was that experts would often have no cornerstone cases for the first few rules they defined, until they felt comfortable enough to start accepting cases. While obviously potentially detrimental, this actually had no effect overall as the first few rules defined were quite general and distinct, and there were no conflicts to be found in the first few cases classified.

3.3.3.2 Statistical Tools

The experts involved in this portion of the study did not themselves use the statistical tools available to them, there being some concern that it would be too time consuming to examine. As was described, the statistical tools were used by the administrator when consolidating the knowledge bases by answering the experts' queries about how their differing rules compared over the data. This application of the tools proved very effective in resolving the experts' knowledge conflicts, quantifying their differences and examining the impact of potential changes in definition. The results were found very quickly, largely being a matter of selecting the appropriate classifications and comparing their results, and entering new rule criteria that the experts requested to examine the impact. It is expected that if the conflict were identified by the experts themselves in a collaborative setting, and the experts were comfortable with examining the statistics themselves, the matter could have been resolved between the experts without any intervention. This remains to a more specific study to determine however.

3.3.4 Impact of Implementation Restrictions

3.3.4.1 Rule Creation

The effect of not allowing disjuncts in rule creation was only noticeable in four instances, three of which were grading severities of the same classification. The fourth was a classification that commonly occurred as a compound: the *Fixed* portion of the many *Fixed obstruction* classifications. Hence although it only seemed to cause a problem in one instance it probably should have caused many more, with the experts simply not encountering other problem cases. This was not overly problematic in the consolidated knowledge base as all of the *Fixed* classifications were consolidated into a single class. The lack of disjuncts did prove to be an effective way of keeping rules simple and easy to compare however, and kept processing time to a minimum for recording classifications and determining statistics for validation; all of which was helpful in ensuring there were no conflicts between experts and in comparisons of expert practices.

3.3.4.2 Rule Conditions

The effect of only having simple conditions, and only allowing the comparison of two attributes via adding a new attribute, is difficult to measure as there is no direct comparison to be made in this domain. There is assumed to be no negative impact, as the experts expressed no confusion over defining rules in this format, nor were there any examples of an expert attempting to use a different format. In fact, every attribute used in a rule condition, by any of the three experts, was a compound attribute derived from at least two other attributes. The time taken to define rules also indicates no difficulty with this rule implementation style.

A further examination of the rule conditions shows that of the 115 attributes available, only 20 (17%) were used by the three experts in rules; 18 in the individual knowledge base and 13 in the collaborative; and as mentioned all were compound attributes, derived from other available attributes. If this is taken into account and they are broken into their distinct components, this gives an actual figure of 50 attributes used (44%). While still seemingly low this is not especially unexpected: extra attributes are recorded from the same testing procedures as those more commonly used, and hence cost nothing extra to calculate but are included for completeness. Each of these extra attributes also has an associated predicted reference value, leading to a large number of extraneous attributes.

3.3.5 Rule-Based Thinking

There is a concern with this study, which was also raised in the EMCRDR study (Ling, 2006), that there may have been a detrimental shift of focus by the experts from case-based to rule-based thinking. In both studies the experts exhibited a more than desirable focus on the rules being entered rather than the cases being classified: each expert would be consciously trying to define correct rules rather than correctly classify cases, which can lead to a series of detrimental effects relating to the change from an evidence-based, cased based reasoning approach, to a more simple expression of conscious knowledge.

It appears that some of the modifications implemented here did have this effect. The clearest evidence that such a shift occurred was the tendency from every expert to focus on making individual rules correct, at the expense of classifying cases completely before moving on to a different case. Each expert chose at some time to

move on to working on another case, or go back to a previous case and make changes, without completing the classifications for the current case. This is highlighted by the number of cornerstone cases which received further classifications after the experts had already accepted them as correct: 24 in the collaborative knowledge base and 42 in the independent. This includes experts re-examining and modifying the classifications for individual cases. These numbers are further bolstered by the number of cases that were viewed without being completed: 41 in the collaborative knowledge base and 33 in the independent. While none of these incidents can be strictly interpreted as an expert switching between cases in order to complete an individual rule, the frequency of movement between incomplete cases suggests that the experts have a tendency to consider the cases as a dataset to be used to define rules, rather than as a series of individual cases for incremental classification.

This is not a surprising result: given that the ultimate goal of this project is to learn new information from the dataset, the cases have deliberately been presented as part of a dataset for defining rules. It is also important to note that the effects of this shift are both positive and negative. As described previously, the extra validation allowed by the dataset provides benefits to the knowledge acquisition, likely benefitting in the accurate acquisition of tacit knowledge; and a rule-centric approach improves the speed of knowledge acquisition in the early stages of the process, as the expert can simply define the rules they are most familiar with. However, the risks are that as the cases are not completed the validation will be less likely to be able to assist; and if cases are only examined to a shallow level before continuing it is very unlikely to reveal any tacit knowledge.

As none of the evidence listed here is uncommon in RDR knowledge acquisition, nor does this mode of thinking about the process seem unlikely, it is surmised that this rule-centric thinking is probably reasonably common in RDR knowledge acquisition; though perhaps not to the extent that is apparent in this study. While it may occur in any RDR knowledge acquisition, it seems reasonable to assume that the effects are fairly limited, given the success of the RDR method in general. Nevertheless care should be taken to avoid compromising the evidence-based nature of the knowledge acquisition and validation wherever possible.

3.3.5.1 Misunderstanding the rule structure

Many of the biggest problems in the knowledge acquisition, and especially in the knowledge consolidation, occurred due to misunderstandings of how the knowledge base structure functioned. Most experts expressed confusion at least once about exactly how the rule they were entering would affect the outcomes of the knowledge base. In almost every case the experts' worries resulted from over-thinking the problem, and a lack of understanding of how the rule structure worked. It is possible that these issues were more prevalent due to the exhibited rule-based thinking: as the experts were concentrating on defining a knowledge base for a dataset, rather than on classifying cases, they may have become overly concerned about how to define the most effective rules rather than in classifying each case correctly. This can have a detrimental effect on the knowledge acquisition, by reducing the likelihood of acquiring tacit knowledge, and by causing confusion as the expert tries to understand the knowledge base structure and, as was seen in this study, makes mistakes in rule definition through that confusion.

These misunderstandings occurred quite frequently with experts unsure about the optimal way to resolve a misclassification, trying to make sweeping changes to the knowledge base outcomes with a single rule. In all instances when any such issue arose, the administrator made the recommendation to the expert that they should focus on the case they were working with – simply make sure that all classifications are correct for the current case. Once correct, move on to the next case regardless; the system will ask you to solve any error when it actually becomes a problem. If the current classifications are not correct, then resolve them following the usual steps: select the erroneous classification and choose the *Change conclusion* option; then describe why the current classification is wrong. This advice was accepted by each expert when given and resolved the immediate problem, but the issue still recurred afterwards, and would be expected to continue to occur if the knowledge acquisition continued.

One of the most common problems causing such uncertainty was the expert retrospectively viewing a rule and deciding that the rule had been entered incorrectly – a typical example is a desire to tighten a rule by changing a condition from *less than 70* to *less than 60*. The correct way of resolving this issue in traditional MCRDR is to add an exception or stopping rule, specifying that if the

value is greater than or equal to 60 then the alternate classification, or no classification, should apply. However, as has been seen by the way in which experts attempted to correct rules, when considering the knowledge base in terms of the rules that it contains this is not an intuitive step to make. Intuitively, the rule condition should just be changed from 70 to 60, and the condition the expert expects to be adding is *less than 60* rather than *greater than or equal to 60*. The examination of the classifications which used multiple rules showed how, after realising that they could not simply change an existing rule, the experts would not reach the correct solution and would attempt to add an exception rule with the same classification hoping that it would override the current (incorrect) rule.

In this instance, following the advice to focus on the case and its classifications rather than the rules, their action should be to do nothing: the case has the correct classifications already, so they should move on to the next case and wait until the error presents itself by incorrectly classifying a new case. However this is a difficult step to take when the expert already knows that the problem exists, particularly when they understand that their focus is to be training the system with their knowledge: with accompanying concerns that they might forget about the error, the system might misrepresent their knowledge to other experts, and that this error might be the cause of more significant problems in the future. This problem is much more significant when the expert is focussed on defining correct rules, as the experts often appeared to be in this study. This is unfortunately a problem with using MCRDR to acquire knowledge for the sake of acquiring knowledge; while MCRDR acquires knowledge well through routine use, where experts are focussed on completing each case correctly and not at all concerned with the structure of the knowledge base, it performs less well otherwise.

3.3.5.2 Irreparable Mistakes

On the problem of correcting a loosely-defined rule, it is also possible that the mistake cannot be corrected under the MCRDR method. It is quite conceivable that there may not be another case in the dataset which is covered by the incorrect portion of the rule, in which circumstance the knowledge base is wrong, and known to be wrong, but cannot be corrected. To illustrate: assuming that there is a case with a Temperature of -10 degrees. A rule is defined of the form $Temperature \leq 5$

→ *Freezing*, where $Temperature \leq 5$ is the condition and *Freezing* the classification. After entering this rule, the expert realises that this is incorrect: the condition should state $Temperature \leq 0$. To resolve this there should be a stopping rule added with the condition $Temperature > 0$. However, this cannot be added under the context of the existing case: the condition $Temperature > 0$ does not apply to this case, and will be rejected. At any rate, the case should have the classification *Freezing* and so removing the classification under the context of the current case would be wrong. Instead, the expert would continue going through cases until they find a case which has a Temperature between 0 and 5, which would be classified incorrectly by the current rule, and provide the opportunity to add the stopping rule. The issue arises when there are no cases with a Temperature between 0 and 5: the rule cannot be corrected, as there is simply no context to provide justification that the rule should be changed.

There are various solutions to this problem. Perhaps the most obvious solution is to allow the expert to edit rules, as was used in the EMCRDR study. This however goes against the general RDR philosophy that knowledge should never be removed, only added to, based on the assumption that knowledge that is entered is correct in the context it was entered in, and therefore should remain as long as that context (i.e. the case it was based on) is believed to be true. This assumption is valid in this situation, and helps to ensure that knowledge is only entered when there is supporting evidence (a case demonstrating the principle represented by the rule). However the problem in this example is that the knowledge entered is correct, but not correct enough.

The common way to correct the mistake, and the only solution that strictly adheres to the philosophy of never removing knowledge from the knowledge base, is to add a stopping rule with no conditions to the incorrect rule, under the context of the original or a similar case. This will cause the classification *Freezing* to never be reached by this rule, as it is always overruled by its stopping exception. Then a new rule is added for the case, stating the desired knowledge – $Temperature \leq 0$ → *Freezing*. While this approach works perfectly well, it can cause the knowledge base to become cluttered with useless rules that will never fire, which has two negative impacts: firstly, it takes up unnecessary processing time, and secondly, these rules can make interpreting the knowledge base very confusing when

requesting an explanation for how a classification was reached. The various strategies used to restructure and reduce a knowledge base can be implemented here; however a periodic post-processing, taking the system out of use and restructuring is not ideal as it requires specific and periodic intervention. As mentioned, this is the approach that is taken in typical MCRDR developments, and studies have failed to find significant detrimental effects from this process (Kang, 1996; Suryanto, et al., 2002). This is the method that was used in this study, as a part of the knowledge base consolidation, with the administrator manually correcting these rules where appropriate.

A refinement of the rule editing/removal device used in the EMCRDR study provides another option: the option of ~~"u~~ndoing" the last created rule, rolling the knowledge base back to before the newest rule was added so the expert can add it again more correctly. This would not have solved all of the instances of such problems in this study however, where many of the rules with problems were discovered after a few more cases had been examined and further rules added, so is unlikely to be of much benefit in general knowledge acquisition.

A more extreme solution is to allow the expert to define rules outside the context of a particular case. This however allows any rule to be defined without requiring any evidence, which removes one of the most basic rule validation mechanisms, removes all evidence to support and justify the rules, and negates many of the advantages of the MCRDR system. If cases have been reviewed, or are subsequently reviewed, then cornerstone cases may still be identified for such rules.

3.3.6 Interface Issues

Many of the issues encountered, such as implementing rules incorrectly, can be attributable to the experts misunderstanding the system or the interface. It was noted that of the 15 experts who at various times tested the system and defined rules, those with more obvious familiarity with working electronically encountered less problems. In particular those experts who were familiar with working in an online environment (specifically web forms and related technologies) had very little trouble using the system as directed. Indeed the student users who participated in the knowledge comparison study, described in section 5.3.2, showed remarkable aptitude in identifying within seconds what options were available to them, and to

be able to clearly identify, understand, and follow prompts. In contrast, the more experienced experts were generally much more hesitant in using the system, had difficulty identifying options, and generally had difficulty following the interface. It is difficult to ascertain exactly which errors were caused by these interface problems; however they would seem to have had some effect on increasing the number of misunderstandings and definition problems. This is a failing of the lack of training provided to the experts in how to use the system, where more practical instruction was needed rather than online tutorials and documentation. This is more a point of consideration for any future development rather than a significant problem, as although the errors slowed knowledge acquisition and made consolidation more difficult, they do not appear to have significantly affected the end result.

3.3.7 Multiple Experts

3.3.7.1 Identified Errors

Listing the creator of a rule seemed effective in reducing the number of conflicts: it was observed that experts would be more cautious and thorough about changing rules when they were aware that it was another expert's input, and they were not just correcting their own previous mistake. It was also observed that the experts were more comfortable changing the rules entered by the administrator than rule entered by another expert, especially when there was a feeling that the other expert would probably know better than themselves. However, it is assumed that this was successful partially because each participant was at least acquainted with each other expert, and knew of their qualifications and experience. For a larger scale study including many experts it may be important to, with participant consent, list each expert's credentials so that other experts can check who is disagreeing with them, and to perhaps include a simple facility for communication between the experts. This would keep the system open to many experts of many backgrounds. It would also be an interesting study to ascertain what differences occur between displaying expert credentials and not: for a knowledge discovery system it may well be beneficial to attempt to convince experts to not dismiss any rule without examination of the data, by not letting them see who created the rule. This would, however, need to be balanced with the range of experience and knowledge of the

participating experts, to avoid having experts take too much time needlessly questioning domain principles, and would require an effective conflict resolution strategy.

3.3.7.2 Identified Conflicts

While there were multiple strategies in place to resolve conflicts in expert opinion, in practice it was found to require more direct intervention. The case highlighting approach, showing cases that had been disagreed with in red, proved unsuccessful due to the difficulties in organising the timely participation of multiple experts. Due to the typical restrictions on expert availability, experts preferred to work with the system in discrete time periods, trying to consolidate their interactions into as few sessions as possible. The busy and conflicting schedules of each expert further meant that these sessions were often quite separated, resulting in there being very little overlap in time between the inputs of one expert and another. Hence, it became very unlikely that an expert would ever, without direct prompting, access the system and see cases which had been disagreed with. With the possibility of experts noticing past conflicts effectively removed, it became reliant on the expert who made the new rule to contact the other expert and correct them or open discussions, which never occurred: given that the experts were busy, they unsurprisingly were not interested in taking the time to start a dialogue over every potential difference; each of which may only represent an input error. There was generally an attitude that the other expert had either only made a minor oversight, or were simply mistaken and that they would most likely realise their mistake if they ever looked at it again; further, that now that the system was correct (in their view), why should it need any further discussion? With these simple justifications and a busy schedule, it became clear that these methods of conflict identification would not work, as the likely outcome would be that only one party would be aware of any disagreement, and would probably pay little attention to it, considering it fixed.

Given the lack of success of expert-initiated methods, the conflict resolution fell to the other alternative: which was for the administrator of the system to resolve the conflicts by contacting the experts involved and initiating a discussion. This was a far from ideal situation as the administrator lacked the expertise to be able to interpret the classifications accurately. This combined with the different phrasings

and terminology used, and different levels of detail that experts used in their classifications, made identifying true conflicts and communicating them to the experts a more difficult and inefficient task than it may have been. Nevertheless, of the 5 conflicts identified, all were easily resolved by email discussions within a few days of being identified. This approach is considered successful however, as it kept required expert time to a minimum, which is still the bottleneck and major problem faced in such work.

There were minor miscommunications due to misunderstandings of the domain from the non-expert administrator. It is suggested that in a larger scale project this may need to be resolved by having someone with sufficient expertise in the area act as mediator; or, by ensuring sufficient commitment from the experts involved that they would be able to regularly access the system and review cases, in an overlapping time period. It is expected that in a commercial setting with definite outcomes for the experts involved, especially if they are paid to participate, that either of these would be a viable possibility.

A third option of conflict resolution was considered for this study, whereby whenever a conflict arose from an expert changing a previously accepted case, the system would generate an email to the expert who made the initial classifications thereby automatically initiating discussions. It became apparent that this would be impractical for the experts involved as their schedules for interacting with the system were widely deviant, and they would not welcome unsolicited emails, particularly if many were minor errors that required no further action. Due to the lack of verification in this process and the potential for a large number of emails to be generated, this is not recommended as the best solution; particularly with the possibility of frustrating the experts who are, in any knowledge acquisition venture, the most valuable and important resource.

3.3.8 Classifications as Rule Conditions

It was noted at many stages throughout the knowledge acquisition process, with many different experts, that one of the initial instinctive responses for experts was to define new classifications in terms of previous classifications; or to put it another way, to use existing rule conclusions as new rule conditions. This behaviour generally disappeared once the expert was informed that this was not possible,

requiring only a minor mental shift to always think about classifications separately. However the desire would still occasionally manifest in an expert attempting to define sub-classifications by adding exception rules, usually without a complete understanding of what the end result would be.

3.4 Conclusions

The methods described here have resulted in a functional expert system for lung function interpretation: a collection of reproducible expertise on how to classify lung function cases. This knowledge was successfully compiled from the acquired knowledge of multiple experts, both as a collaborative acquisition effort and as a post-acquisition consolidation. The use of an extensive dataset of lung function cases provided additional evidence-based validation of the knowledge provided, which was especially useful in the comparison of expert knowledge and the resolution of knowledge conflicts. Additionally, although the results expressed here show an MCRDR expert system that conforms to the usual standards of such systems, an analysis of the dataset suggests that the system's knowledge may be incomplete.

It is important to note that the methods have several limitations. Providing the dataset for extra rule validation invoked a slight shift away from a case-based focus to a rule-based focus, confusing and slowing the experts in the knowledge acquisition process. Whether the benefits provided by these modifications outweigh the detriments is unclear, but should also be considered in light of the applications of those modifications presented in the following chapters.

Both the processes of having multiple experts work within one knowledge base, and of having multiple individual knowledge bases compared and combined seemed to function effectively based on the results that were available. The collaborative knowledge base seems to be a more efficient method of consolidation, with much less effort required by an administrator, but results from this study are far from conclusive on this point. The post-acquisition comparison and consolidation provided much greater benefit in identifying differences of opinion and resolving them, to both experts' satisfaction and education. Again, however, a comparison between the two is inconclusive.

The study does show that there are significant differences between experts' practices in the lung function domain: even after having removed terminology differences between experts, 63.1% of the dataset received different classifications, some of these conflicts even occurring between classifications reached by the same expert at different times. This would suggest that the consistency that could be provided by an expert system such as the one described in this thesis could be greatly beneficial. Similarly the potential for collaboration to develop agreed upon standards is a major benefit.

The produced expert system can be of benefit to the lung function domain, assisting experts in interpreting cases; of more relevance to this study however, it also provides a store of knowledge that can assist in knowledge discovery and data mining, as will be explored in the next chapter.

Chapter 4 Knowledge Discovery and Development

4.1 Introduction

In this age of ubiquitous electronic data logging, transfer and storage, enormous amounts of archived data are being created and added to by every action we make and decision we take. A recent study has estimated that 295 exabytes ($2.95x10^{20}$ bytes) of data were stored in 2007, following a 23% increase per year since 1987 (Hilbert & López, 2011). If analysed much of this data could reveal patterns, representative of recurring events or trends, which could be used to predict future events and assist in decision making in a huge range of fields (Witten & Frank, 2005). Such results can provide not only monetary benefits in areas such as identifying business trends and improving working efficiency, but also in critical areas such as environmental prediction or disaster prevention and management (D. Zhang, et al., 2002).

Health and medical data is being electronically archived as much as any other form of data (Cios & Moore, 2002a; Prather, et al., 1997; Steinberg, Wang, Ford, & Makedon, 2008). It was estimated in 2002 that three quarters of a billion people had medical data recorded in electronic form in North America, Europe, and Asia (Cios & Moore, 2002a). The analysis of medical data presents unique challenges, but can provide unique benefits: if done successfully, it has the potential to provide improvements in health care for the population, and improvements in quality of life for the individual (Cios & Moore, 2002a; Roddick, et al., 2003).

Knowledge discovery is the process of analysing archived data in order to find new knowledge (Goebel & Gruenwald, 1999). A knowledge discovery process is not simply an automated mathematical comparison or logical inference however. In order to discover truly useful and new knowledge, a level of existing knowledge about the data must be identified and incorporated, and the results of the data analysis must be examined and interpreted by a human expert (Fayyad, et al., 1996a; Liu, et al., 1997; Piatetsky-Shapiro & Matheus, 1994; Pohle, 2003). This is especially true in medical knowledge discovery, where there is already a vast

amount of complex existing knowledge. The results of knowledge discovery from medical data also require specific and extensive expertise to interpret, and the validity of the discovered results are critical (Cios & Moore, 2002a; Roddick, et al., 2003).

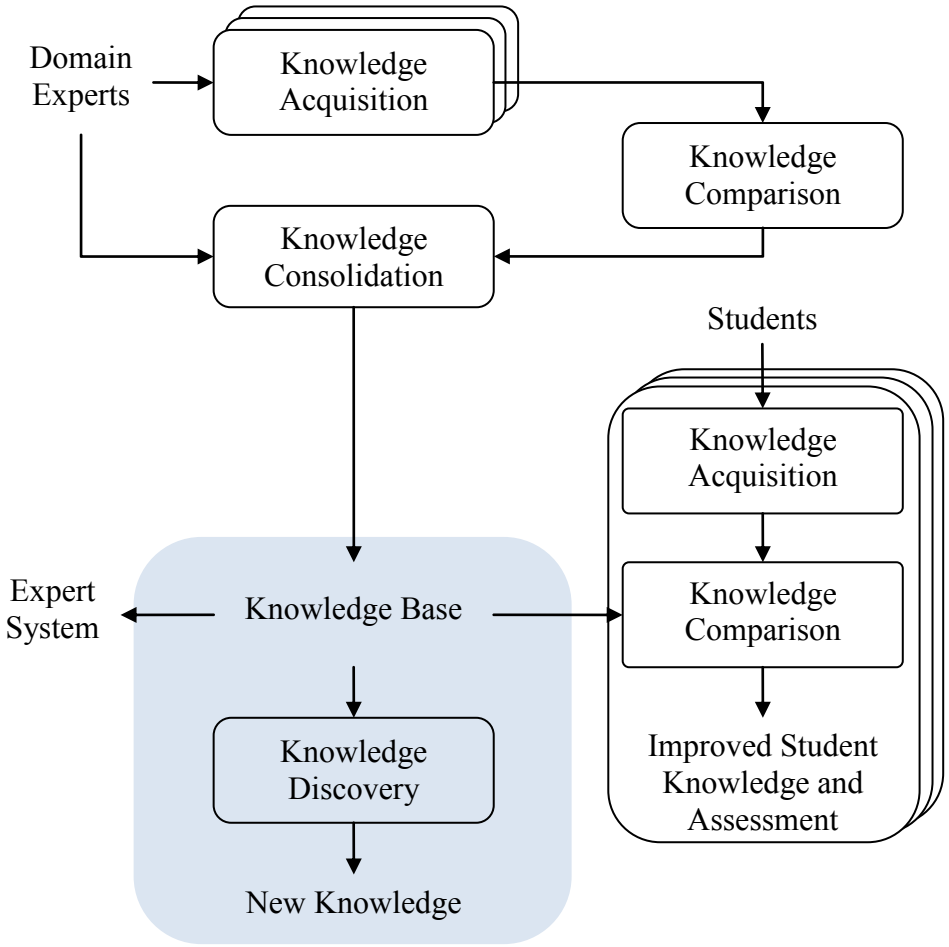


Figure 4-1: The methods presented in this thesis; the highlighted section shows the components presented in Chapter 4

This chapter will present a new approach to knowledge discovery in complex domains, which can effectively incorporate existing knowledge in data and results analysis, and can integrate newly discovered knowledge into a central knowledge base. This is achieved by involving the user at every stage of the process, and using knowledge acquisition techniques to reduce the costs of expert involvement. The method is tested in the field of lung function. The knowledge base and knowledge interface described in section 3.2.2 are used to provide context for data exploration, while the addition of data mining techniques provides assistance in discovering new

relationships or patterns. These patterns can then be interpreted by the user and formalised as new knowledge, which is automatically incorporated into the knowledge base. The knowledge discovery component of the study is highlighted in Figure 4-1.

4.2 Methodology

4.2.1 Structure

The method presented in this study allows a user to explore a dataset to discover new relationships and patterns, with the benefit of having identified existing patterns. The knowledge discovery method consists of a set of distinct components: a database of cases, a knowledge base of rules to classify those cases, a composite of functions for calculating statistics and mining interesting relationships in the data, and an interface for a user to explore these elements and direct the functions. The structure of these components is summarised in Figure 4-2. The database, the knowledge base, and the process used to build the knowledge have been described in section 3.2; this section will describe the exploratory analysis functions, and the process of discovering new knowledge via statistical calculation and data mining.

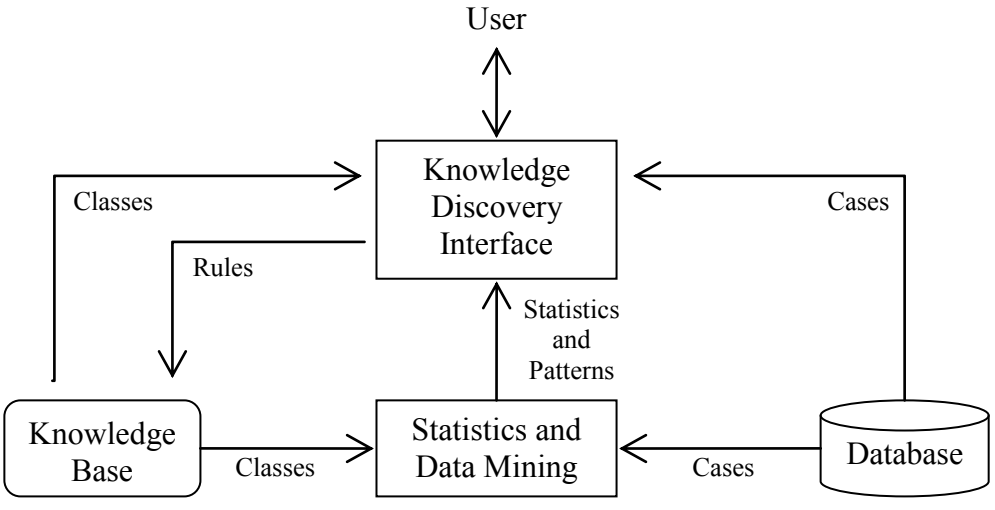


Figure 4-2: Simplified structure of the knowledge discovery process

4.2.2 Data Analysis

The data analysis functions include simple data mining and statistical calculations, which can provide additional data and suggest further links within the database and the knowledge base. These are incorporated into the online knowledge acquisition and expert system described in Chapter 3 through a dedicated data search and analysis interface. By incorporating this into the RDR process, an environment is created for exploring trends in the data and testing hypotheses about why those trends exist: the knowledge definition options provided by the RDR system can be used to identify data of interest, and the data mining functions present any interesting relationships for that data.

These statistical and data mining components are presented in two different sections of the system: a case search and statistics page, and as was briefly described in the preceding chapter (section 3.2.2), a section incorporated into the rule definition page providing additional validation to the knowledge acquisition system.

4.2.2.1 Case Set Statistics

The primary interface for the knowledge discovery functions is the search and statistics page. The page contains a web form allowing the definition of class and attribute criteria to define a subset of cases (shown in Figure 4-3). The form is very similar to the rule definition interface, with the notable exception that the existence or absence of a class can be included as a condition. These allow the user to identify specific subsets of cases to work with or examine; for example, all cases who display attribute *x* and not *y*, or have *Class A*. The inclusion of a class as a condition uses the maintained case-class-rule records to identify which cases should be included in the set, for reasons of efficiency only: if the class conditions were converted to their constituent rules, and those rules converted to their constituent attribute conditions, the resultant case set would be identical. In essence, by including a predefined class as a search condition, the user is specifying that at least one of the rules which currently lead to that conclusion must be true. Disjuncts are also provided as an option for all of these conditions.

The page then presents a summary of statistics for that set of cases, compared against statistics for the entire dataset, including highlighting any unexpected relationships identified through the data mining functions. The use of the entire

dataset for comparison provides expected values for the attributes and expected frequencies for the classifications, under the assumption that the dataset is representative of the wider domain. However the user may also choose to select another set to compare against other than the entire dataset, if there is some more specific comparison to be made. There is finally an option for the user to transfer their existing search conditions to the rule definition screen, to define a classification for the current set. These calculations will be described in more detail in the following two sections.

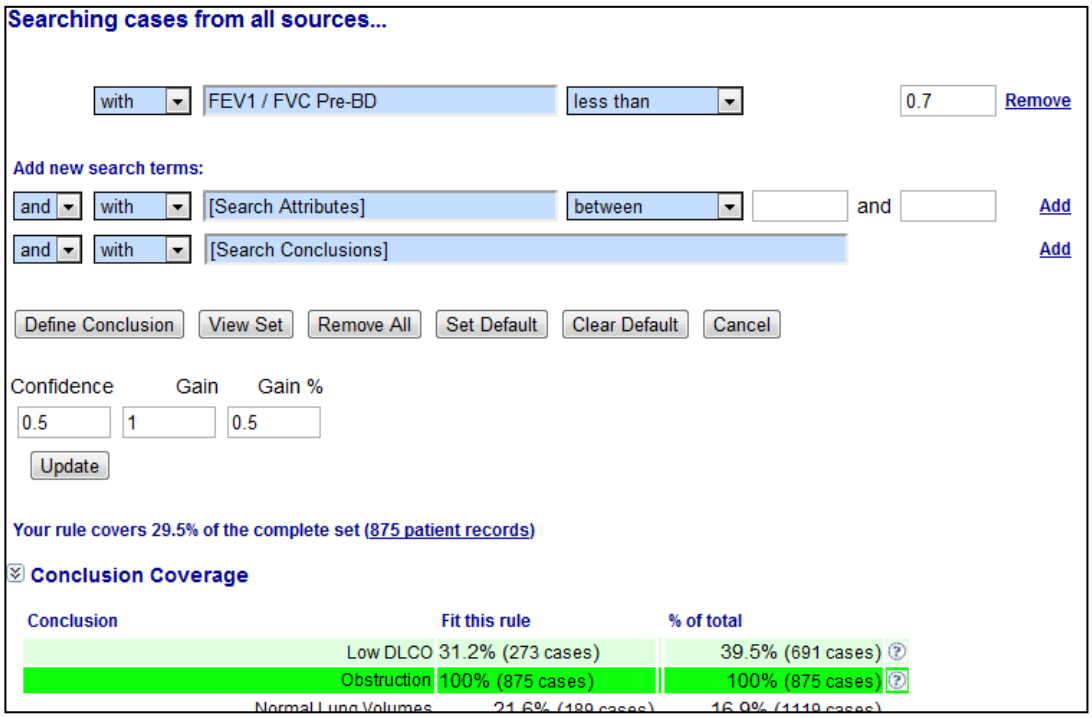


Figure 4-3: Partial screenshot of the case set search and statistics screen

4.2.2.2 Rule Statistics

The rule statistics section contains the same information as the search page statistics, but presented concurrently with the rule definition form. The statistics it presents relate to the rule as defined by the current conditions and currently selected class.

In early versions of the system there was a clear distinction between the combined search and data analysis section, and the rule definition section, with statistics provided only for the search and case set component. As work progressed it became more apparent that the two were closely related, until by the end of development the two could have been integrated entirely into one function. The only distinction

between the search screen and the rule definition screen is that the search screen allows the inclusion of conditions based on classes; and the rule screen requires that the rule be given a classification, which it uses to perform MCRDR validation against cornerstone cases. Neither of these present a distinct problem: the inclusion of a class for the current set of conditions is trivial, as it can simply be any placeholder until the user selects a name; and the inclusion of an enhancement such as Recursive RDR or MCRRR would allow class conditions to be legitimate rule conditions. This highlights the closeness of the processes of knowledge acquisition and knowledge discovery: one, applying an expert's knowledge to a set of data in order to formalise that knowledge; and the other, analysing a set of data using expert knowledge, in order to formalise some new knowledge. While this is a generalisation, and the specifics of how the data is examined and how knowledge applied are different, this nevertheless shows the similarity between the two processes.

4.2.2.3 Statistics and Measurements

The measures used were primarily taken from association rule mining, information theory and probability measurements, given their applicability to exploratory data analysis and data mining in general (Creighton & Hanash, 2003; Lenca, et al., 2006; Marinica & Guillet, 2009). The measures were chosen for their ease of calculation, such that the system can be sufficiently responsive with a large database in an online environment, and for their simplicity, such that the experts could understand what was being indicated.

In both the independent search screen and the rule definition statistics segment, the presented information is divided into three sections: class coverage, attribute statistics, and a summary of the 10 most correlated attributes. Each section is expandable on the screen to avoid unnecessary clutter.

Class Statistics

The class section displays each of the classes that are present in the case set currently being considered. For each class, the number of cases that have that class is displayed, along with the percentage of the current set which that number constitutes. Displayed for comparison are the number of cases that have each class

in the complete set, and finally the percentage of all members of that class that are in the current set. Each case count has a link for the expert to view and work with that set of cases, and each class has a link for the expert to view the rules that define it in the knowledge base.

To determine the significance of relationship between the current set and each class, a number of measures are calculated based on the principle that the current conditions are the antecedent of an association rule, and the class is the consequent, in the form $(current\ search\ conditions) \rightarrow class$. These measures included the confidence, gain, and Piatetsky-Shapiro gain (hereafter referred to as *p-sgain*). The support was displayed explicitly, for the rule and for each class, so is not included in interestingness calculations. The user has the option to specify three thresholds: a confidence threshold α, between 0 and 1; a gain threshold β, between 0 and *n* (size of the set); and a gain percentage threshold γ. The calculated measures are then checked against the user-variable thresholds to determine the interestingness of the relationship: if the confidence of the rule exceeds the confidence threshold $(\frac{support}{n} > \alpha)$, the confidence is significant. The *p-sgain* measure returns the number of cases that have the class above what would be expected, based on the ratio of that class in the whole dataset. If this measure exceeded the user modifiable threshold β the class was marked as significant. To normalise this measure and allow a more even comparison between large and small classes, an extra measure was defined by dividing the *p-sgain* by the support of the rule. This was then compared to the user-defined threshold γ, and again if it exceeded the gain percentage threshold $(\frac{p-sgain}{n} > \gamma)$, then the class was marked as significant. In the web interface, class statistics were highlighted progressively stronger shades of green the more measures were found to be significant.

Attribute Statistics

Below the class statistics are listed each attribute, grouped into six expandable sections: *Patient Details*, containing attributes such as sex, age, and height; *Common Spirometry* containing those spirometric measurements most commonly used by experts, such as FEV_1, FVC and FEF_{25-75}; *Other Spirometry* containing 27 less commonly used spirometric measurements; *Lung Volumes*; *Gas Transfer*; and

the source of the data. The attributes contained in each of these groups, and the possible values for non-numeric attributes, are listed in Table 4-1.

Patient Details	FEF_{25-75} Pre-BD	FRC
Sex	FEF_{25-75} % Predicted Pre-BD	FRC % Predicted
F	FEF_{25-75} Post-BD	IC
M	FEF_{25-75} % Predicted Post-BD	IC % Predicted
Age	**Spirometry: Other**	ERV
Height	FEV_1 Δ after BD	ERV % Predicted
Weight	FEV_1 Δ % Predicted	SVC Pre-BD
BMI	FEV_1 / SVC Pre-BD	SVC Post-BD
Smoker	FEV_1 / PEF Pre-BD	SVC % Predicted Pre-BD
yes	FEV_1 / PEF Post-BD	RV / TLC
no	FVC Δ after BD	RV / TLC % Predicted
ex-smoker	FEV_3 Pre-BD	V_A / TLC
missing	FEV_3 % Predicted Pre-BD	R_{aw}
Pack Years	FEV_3 Post-BD	**Gas Transfer**
Spirometry: Common	FEV_3 % Predicted Post-BD	D_LCO (Hb corrected)
FEV_1 Pre-BD	FEV_3 / FVC Pre-BD	D_LCO (Hb corrected) % Predicted
FEV_1 % Predicted Pre-BD	FEV_3 / FVC Post-BD	
FEV_1 Post-BD	FET Pre-BD	V_A
FEV_1 % Predicted Post-BD	FET Post-BD	V_A % Predicted
FEV_1 % Δ after BD	FET Δ after BD	D_LCO / V_A (Hb corrected)
FVC Pre-BD	FIF_{50} Pre-BD	D_LCO/V_A % Predicted (Hb corrected)
FVC % Predicted Pre-BD	FIF_{50} Post-BD	
FVC Post-BD	FEF_{50} Pre-BD	D_LCO (uncorrected)
FVC % Predicted Post-BD	FEF_{50} Post-BD	D_LCO % Predicted (uncorr)
FVC % Δ	FIF_{50} / FEF_{50} Pre-BD	D_LCO / V_A (uncorrected)
FEV_1 / FVC Pre-BD	FIF_{50} / FEF_{50} Post-BD	D_LCO / V_A % Predicted (uncorrected)
FEV_1/FVC % Predicted Pre-BD	PIF Pre-BD	
FEV_1 / FVC Post-BD	PIF Post-BD	TLC - V_A
FEV_1 / FVC Δ	FIF_{50} / PIF	COHb
PEF Pre-BD	FIF_{50} / PIF Post-BD	Haemoglobin (Hb)
PEF % Predicted Pre-BD	SVC / FVC Pre-BD	D_LCO % Predicted / D_LCO / V_A % Predicted
PEF Post-BD	IVC	
PEF % Predicted Post-BD	**Lung Volumes**	**Other**
FEF_{25} Pre-BD	TLC	*Source*
FEF_{25} Post-BD	TLC % Predicted	AH
FEF_{75} Pre-BD	RV	RHH
FEF_{75} Post-BD	RV % Predicted	TAHS

Table 4-1: Statistics are calculated for these lung function attributes

For each of the numeric attributes, whenever statistics were displayed the minimum, maximum, mean, and the standard deviation were listed. For the nominal attributes, each allowable value was listed, along with a count of how frequently that value occurred in the current working set, and for comparison, how frequently it occurred in the overall dataset.

Each attribute is also tested for potential *interestingness*, as defined by an unexpected association to the current set of cases. This was tested through various measures. Firstly, by using the range provided by the minimum and maximum values in the current set as the consequent of an association rule, and the conditions chosen by the user as the antecedent, in the form:$(current\ search\ conditions) \rightarrow (attr_{min} \leq x \leq attr_{max})$. As with the class statistics, the attributes were rated with confidence, gain, and *p-sgain*; for numeric attributes, the *z*-score of the mean for the current set (the difference between the mean of the attribute for the current set and the mean of the attribute for the entire set, divided by the standard deviation), was also calculated. If the difference exceeded the user-defined confidence threshold α, the attribute mean was marked as interesting, and the magnitude of the score provided. Likewise for gain and *p-sgain*, if the calculated values exceeded β or γ, respectively, then the attribute was marked as *interesting* (having some association), and the magnitude of each difference displayed with the attribute. Associated attributes were highlighted green and the magnitude of the significance indicated in simple terms, for example "Cases in this range have the rule conditions 50% more often than expected". The attribute was highlighted a brighter green depending on how many measures found an association.

Nominal attributes also had an information gain calculated for each value to determine if the conditions were a good predictor for that value. For numeric attributes, the information gain calculation was not performed automatically, but on request for each attribute (as the user hovers the mouse cursor). When requested the system uses information gain calculations to find the range which optimally predicts the set defined by the current rule conditions. A screenshot showing a small sample of attributes is provided in Figure 4-4.

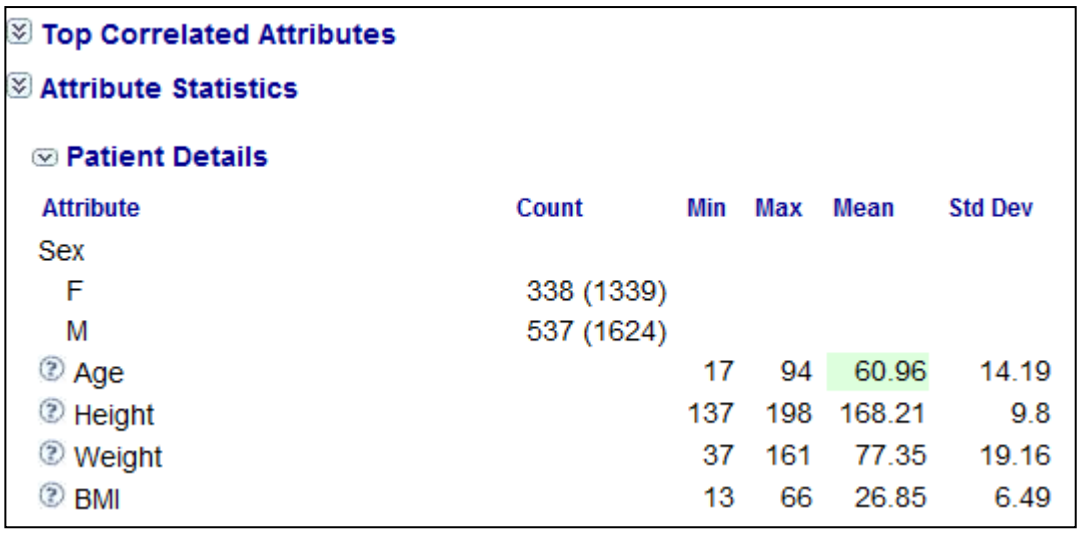


Figure 4-4: A small selection of attribute statistics

Also displayed, separately to the main list of attributes, was a summary of the 10 most associated attributes in order of the normalised *p-sgain* measure, to provide the user a more immediate impression of interesting findings.

4.2.2.4 Knowledge Discovery Process

Figure 4-5 shows the computational process of the knowledge discovery system, from the user defining their set conditions and specifying a comparison set (which defaults to the entire dataset), and finishing with a series of options for the user. Fundamentally the knowledge discovery method is quite similar to the knowledge acquisition method. The user defines search conditions to describe the set of cases they are interested in, in a very similar manner to defining a rule, except that they can also specify conditions requiring the presence or absence of particular classifications. Statistics are calculated for the user-defined set, summarising details such as the ranges and averages for each attribute, and the prevalence of each classification. The same statistics are calculated for the comparison set. The data mining features are then used to compare these two groups of statistics, finding unexpected differences and marking them as potentially interesting, along with the reasons for marking them as such. The user can then further refine their search terms to examine a more specific subset of cases; view more detailed statistics such as finding the best information gain range; view a specific case set defined by a class or attribute range; or, if they believe that the currently selected set has some

property worth recording, they can define a class to apply to that set with the rule generated as from the search terms specified. The new knowledge described by this rule and classification are then validated against existing cornerstone cases, in order to maintain the validity of the knowledge base, and the new knowledge is immediately available for use; as per the incremental knowledge acquisition. As this study did not implement an MCRRR approach, if the user chooses to define a class for the current set and a class is being used as a condition, then the rule leading to that class is extracted and its conditions added to the attribute conditions already defined.

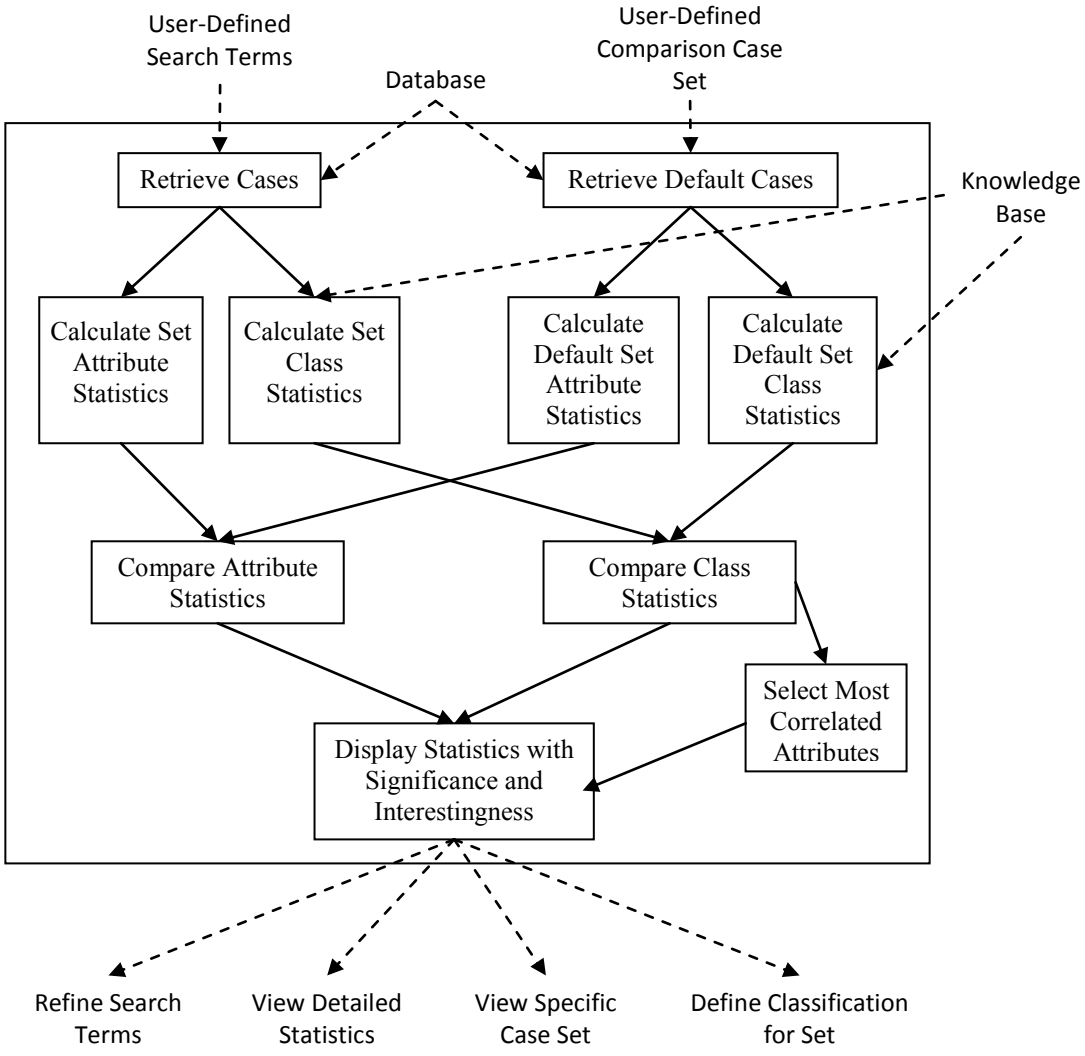


Figure 4-5: Computational Process of the Exploratory Analysis Component

4.2.3 Testing the Method

To test that the method could successfully discover new knowledge from complex data, the system was used to resolve and expand on research questions that were raised in respiratory literature, or that were suggested by lung function experts as interesting topics to consider and answer through data analysis. To this end a series of existing respiratory studies have been examined, and their hypotheses tested within this framework. Respiratory experts were consulted throughout in order to ensure accuracy and a correct understanding, and to interpret results where necessary.

4.2.3.1 Clinical Studies

Study 1

A pertinent study is Agahi's work from 2007, which sought to make similar use of archived data in the examination of three clinical questions in lung function. The study by Agahi used the same dataset as the EMCRDR study of 484 respiratory cases. However, the classifications provided by the RDR classification system were not used in the analysis, with all evaluation carried out manually with Microsoft Access (Agahi, 2007).

Clinical Question 1: Describe the distribution and pattern of lung function of subjects who met the FEV_1 and/or American Thoracic Society positive reversibility criteria.

Assessing whether a patient demonstrates a significant positive bronchodilator response is a critical factor in differentiating asthma from Chronic Obstructive Pulmonary Disease (COPD) (Agahi, 2007). A significant response is indicative of asthma, whereas a patient with COPD will exhibit a minor response or no response at all (Bleecker, 2004; Meneely, Renzetti, Steele, Wyatt, & Harris, 1962); although some studies have questioned whether the degree of reversibility is as distinguishing as previously thought (Bleecker, 2004; Burrows, Bloom, Traver, & Cline, 1987). A substantial component of this problem may be that there is little agreement on what constitutes a significant (positive) bronchodilator response, with laboratories using different definitions. The change in FEV_1 (FEV_1 Δ) is the attribute most commonly used to define reversibility, with either the absolute

change, percentage change, or change in percentage of the predicted value (Agahi, 2007; Borg, Reid, Walters, & Johns, 2004; Jenkins & Young, 2004). Various other measures have been used, such as the FEV_1/FVC ratio or the $FEF_{25-75\%}$, but are not recommended due to being misleading for different FVC values (Agahi, 2007; American Thoracic Society, 1991). The American Thoracic Society (ATS) and the European Respiratory Society (ERS) define a positive bronchodilator response as an increase post-bronchodilator of at least 12% and 0.2L for either FEV_1 or FVC (American Thoracic Society, 1991).

2007 Study Results

Agahi described that 117 (24%) of the 485 cases met the ATS/ERS criteria for reversibility, with 84 cases matching FEV_1 criteria and 65 matching the FVC criteria. Agahi also stated that 31 cases satisfied both FEV_1 and FVC criteria[5]. Also examined was the number of cases that matched the various sub-criteria of the ATS/ERS definition for reversibility, although no conclusions were reached about this data. These findings are summarised in Table 4-2 and Table 4-3.

FEV_1 $\Delta \geq 12\%$ and FEV_1 $\Delta \geq 0.2L$	FVC $\Delta \geq 12\%$ and FVC $\Delta \geq 0.2L$	Both
84 (17.3%)	65 (13.4%)	31 (6.4%)

Table 4-2: Numbers of subjects, out of 485, matching different ATS/ERS reversibility criteria in Agahi's study (Agahi, 2007)

	Cases	% with FVC response	% with FEV_1 response
FEV_1 response	84	36.9%	100%
FVC response	65	100%	47.7%
FEV_1 or FVC	117	55.6%	71.8%

Table 4-3: Ratios of subjects in Agahi's study with different reversibility criteria (Agahi, 2007)

Agahi performed one more detailed analysis for the data: "~~*In*~~ *subjects who met the ATS/ERS criteria is the correlation with D_LCO stronger in the FVC responders than FEV_1 responders?*". It was reported for this question that FEV_1 showed a stronger

[5] These numbers are out by 1, but this has little impact on the findings.

correlation to D_LCO than FVC, with an R^2 value of 0.196 compared to 0.001; and also that the mean D_LCO was higher for FEV_1 respondents than FVC respondents. This was said to suggest an association between FEV1 response and parenchymal function (Agahi, 2007).

Testing Procedures

As compared to the analysis performed by Agahi, the options provided by this system in incorporating existing knowledge, testing new knowledge, and identifying relevant correlations allow a more efficient and detailed analysis.

The analysis began by defining separate classifications for FEV_1 and FVC reversibility, with the rules [*FEV_1 change* ≥ 0.2L] AND [*FEV_1 change* ≥ 12%] → *FEV1 Reversibility*, and [*FVC change* ≥ 0.2L] AND [*FVC change* ≥ 12%] → *FVC Reversibility*.

New Results

The system showed that 7.9% of cases (235) displayed the class *FEV_1 Reversibility*. Of these cases, 28.5% (67) also displayed the class *FVC Reversibility*, a lower ratio than the 36.9% found in Agahi's study. 159 cases (5.4%) showed only *FVC Reversibility*; with 42.1% also showing a significant FEV_1 response. This gives 327 cases matching either of the criteria. These numbers are summarised in Table 4-4 and Table 4-5.

FEV_1 Reversibility	***FVC Reversibility***	**Both**
235 (7.9%)	159 (5.4%)	67 (2.3%)

Table 4-4: Number of subjects, out of 2963, matching different ATS/ERS reversibility criteria in this study

	Cases	**% with *FVC Reversibility***	**% with *FEV_1 Reversibility***
FEV_1 Reversibility	235	28.5%	100%
FVC Reversibility	159	100%	42.1%
FEV_1* or *FVC	327	48.6%	71.8%

Table 4-5: Ratios of subjects in this study with different ATS/ERS reversibility criteria

The percentage of cases overall matching each of the criteria were found to be smaller in this study, most likely because much of the data from this study are healthy patients whereas the data in the Agahi study were all patients referred for respiratory tests. The ratios between groups are roughly equivalent, although a higher ratio of Agahi's subjects with a significant FEV_1 response seem to also have a significant FVC response; but the ratio of FVC responders to FEV_1 responders is surprisingly similar.

When examining the class statistics for the *FEV_1 Reversibility* class, an interesting relationship was immediately apparent: there appeared to be a stronger overlap with cases with the *Low D_LCO* class than with cases that have the *FVC Reversibility* class. The numbers of cases present are summarised in Table 4-6. Of the 235 cases with a significant FEV_1 response (in other words, belonging to the *FEV_1 Reversibility* class), only 67 (28.5%) also had a significant FVC response; whereas 134 (57%) had *Low D_LCO*. However this confidence test is misleading considered by itself. The *p-sgain* measure provided a more educated indication, as illustrated in Table 4-7. As shown, 81% more cases have *FVC Reversibility* than expected, compared to 59% more than expected with *Low D_LCO*. This is repeated, to a weaker extent, with the cases that have a significant FVC response: 45.9% of cases have *Low D_LCO*, with a *p-sgain* of 49.21%. The confidence is still higher between *FVC Reversibility* and *Low D_LCO* than between *FVC* and *FEV_1 Reversibility*, but the *p-sgain* shows a stronger relationship between the reversibility measurements. The identification that an FEV_1 response is correlated to a reduced D_LCO is unsurprising in itself: a reversible FEV_1 result implies that the airflow is reduced, which would likely cause a reduced D_LCO test result. It would be expected that a post-bronchodilator D_LCO test would show an improvement proportional to FEV_1.

	FEV_1 Reversibility	***FVC Reversibility***	***Low D_LCO***
FEV_1 Reversibility	235	67 (28.5%)	134 (57%)
FVC Reversibility	67 (48.6%)	159	73 (45.9%)

Table 4-6: Numbers of cases belonging to various classes (number in parentheses is the confidence measure that the two are related, or, the ratio of the class within the class under consideration)

	FEV_1 Reversibility	***FVC Reversibility***	***Low D_LCO***
FEV_1 Reversibility	-	+54 (+81%)	+79 (+59%)
FVC Reversibility	+54 (+81%)	-	+35 (+49%)

Table 4-7: *p-sgain* scores for cases with *FEV_1 Reversibility* and *FVC Reversibility* (indicates, for a given class, how many more cases have the second class than expected, shown as the number of cases and as a percentage of the class)

The next most significant class relationship displayed was with the *Obstruction* class, as 63% (150) of the cases demonstrating *FEV_1 Reversibility* were also classified as having *Obstruction*. This is an increase of 80 cases (54%) above the expected number (based on the ratio defined by the larger dataset), indicating a strong correlation. For the 159 cases demonstrating significant FVC reversibility (having been classified with *FVC Reversibility*), 108 (67.9%) have *Obstruction*, an increase of 61 cases (56.52%) more than expected. These ratios indicate that the definition of reversibility correlates well with *Obstruction*, as would be expected because an obstructed patient has more potential for improvement.

Table 4-8 summarises the distribution of classes for each definition of reversibility. The numbers suggest that cases displaying a significant FVC response may also have more severe obstruction, but are overall less likely to have a reduced D_LCO, evidence of gas trapping or hyperinflation, and are more likely to in fact have normal lung function.

Class	Expected	*FEV_1 Reversibility*	FEV_1 *p*	*FVC Reversibility*	FVC *p*
Obstruction	875	63.8% (150)	< 0.0001	67.9% (108)	< 0.0001
Mild Obstruction	133	8.5% (20)	< 0.01	3.1% (5)	-
Moderate Obstruction	312	50.6% (119)	< 0.0001	38.4% (61)	< 0.0001
Severe Obstruction	29	4.7% (11)	< 0.0001	8.2% (13)	< 0.0001
Restriction	113	6.8% (16)	< 0.01	6.3% (10)	< 0.05
Hyperinflation	263	17.9% (42)	< 0.0001	13.8% (22)	< 0.05
Gas Trapping	156	23% (54)	< 0.0001	18.9% (30)	< 0.0001
Small Airway Obstruction	81	6% (14)	< 0.01	3.1% (5)	-
Low D_LCO	691	57% (134)	< 0.0001	45.9% (73)	< 0.0001
Normal Ventilatory Function	1007	14.9% (35)	< 0.0001	20.8% (33)	< 0.0001

Table 4-8: Distribution of relevant classes for different ATS/ERS reversibility criteria, with confidence factor for the association, derived from the binomial distribution

Next to be examined were cases that demonstrated significant reversibility together with *Obstruction*, this being a common indicator of asthma (Bleecker, 2004; Meneely, et al., 1962). The numbers are summarised in Table 4-9. Of the cases with *Obstruction* and *FEV_1 Reversibility*, 58.7% of the cases were also classified with *Low DLCO*, a 60.25% increase from the expected ratio. This is a slight but insignificant increase compared to the 57% and 59% found for all *FEV_1 Reversibility* ($p = 0.061$). For the 108 cases with *FVC Reversibility* and *Obstruction*, 49.1% also had *Low D_LCO*, 52.48% more than expected. This appears to be a slightly stronger increase, but is still statistically insignificant when tested with a binomial distribution ($p = 0.062$).

	Cases	*Low D_LCO*	*p-sgain*
FEV_1 Reversibility	235	134 (57%)	59%
FVC Reversibility	159	73 (45.9%)	49%
FEV_1 Reversibility	150	88 (58.7%)	60.25%
Obstruction* and *FVC Reversibility	108	53 (49.1%)	52.48%

Table 4-9: Relationship for *FEV_1/FVC Reversibility* classes, with and without *Obstruction*, to *Low D_LCO*

Most Associated Attributes

For the cases showing significant FEV_1 reversibility, the system indicated a relationship with many attributes, summarised in Table 4-10. Most significantly the percentage change in FVC after bronchodilator administration was highlighted as associated, based on a difference of 1.2 standard deviations from the expected mean of 2.48%, to the actual mean of 11.47%; notably just under the FVC reversibility criteria limit. Likewise the system indicated a relationship to the absolute change in FVC, with a 1.1 standard deviation change from 0.07 expected to 0.35.

As can be seen, the FEV_1/FVC ratio was also highlighted as associated, with a mean of 0.68 taking it below the GOLD threshold for defining obstruction. Subjects with FEV_1 reversibility also showed a significant drop in both pre- and post-bronchodilator PEF, $FEF_{25-75\%}$ % of predicted, and FEF_{50}.

Attribute	Expected Mean	Actual Mean	Std Deviations
FVC % Δ	2.48%	11.47%	1.2
FVC Δ	0.07	0.35	1.1
FEV_1/FVC pre-BD	0.71	0.62	0.7
FEV_1/FVC post-BD	0.74	0.68	0.4
PEF % of predicted pre-BD	93.08%	75.77%	0.8
PEF % of predicted post-BD	95.37%	85.35%	0.5
$FEF_{25-75\%}$ % of pred. pre-BD	77.2%	47.62%	0.9
$FEF_{25-75\%}$ % of pred. post-BD	86.56%	61.84%	0.7
FEF_{50} pre-BD	3.4	1.94	0.9
FEF_{50} post-BD	3.76	2.63	0.7

Table 4-10: Attributes indicated as related to the *FEV_1 Reversibility* class

Cases with FVC reversibility showed associations of varying strength with almost every spirometry measurement. Table A-1 in Appendix A shows these changes. Of note are that the RV showed a large increase from 108.69% of predicted to 136.86%. Similarly the FRC increased from 101.65% to 118.4%. V_A/TLC was also reduced, from 0.86 to 0.78. Lastly diffusing capacity dropped from 84.13% of predicted to 70.48%.

The next analysis considered the secondary question in Agahi's study, ~~"~~*In subjects who met the ATS/ERS criteria is the correlation with D_LCO stronger in the FVC responders than FEV_1 responders?*". As already described in Table 4-9, the system indicated a significant relationship between cases showing either *FEV_1 Reversibility* or *FVC Reversibility* and those showing *Low D_LCO*. The numbers show some relationship, and as with Agahi's findings the relationship with FEV_1 responders is stronger. Examining further, the absolute change of FEV_1 after bronchodilators (FEV_1 Δ) shows a strong Pearson correlation with all D_LCO measurements (R=0.527 for uncorrected D_LCO, p<0.0001), but with no significant correlation between the percentage change of FEV_1 (FEV_1 % Δ) and any D_LCO measurement.

A weaker but still significant correlation is shown between the percentage improvement of FVC and D_LCO measurements, (R=0.295 for uncorrected D_LCO, p<0.0001, with stronger correlations between other D_LCO measurements), although there is no significant correlation shown between absolute change of FVC and D_LCO. The stronger relationship between *FEV_1 Reversibility* and D_LCO, than between *FVC Reversibility* and DLCO, also matches the findings of Agahi.

Comparisons were also made between the set of *Obstructed* cases, and cases having both the *Obstructed* and *FEV_1 Reversibility* classes, in order to examine what else a significant FEV_1 reversibility might be shown to indicate, in the context of obstructed subjects. Some 150 subjects displayed both classes. The class associations are summarised in Table 4-11. The results indicate a statistically significant relationship to *Evidence of Gas Trapping*, *Hyperinflation* and *Low D_LCO*. The indicated attribute association are described in Table 4-12. Expected values appeared for FEV_1 and FVC bronchodilator change. When corrected for haemoglobin, D_LCO showed an increase of 0.4 standard deviations, as did V_A.

Class	**Cases**	***p-sgain***	***p***
Evidence of Gas Trapping	42 (28%)	24 (58%)	< 0.0001
Hyperinflation	34 (22.7%)	16 (47.6%)	< 0.0001
Low D_LCO	88 (58.7%)	41 (46.8%)	< 0.0001

Table 4-11: Classes showing the strongest association to cases with *FEV_1 Reversibility*, for the 150 cases with *Obstruction*

Attribute	Expected Mean	Actual Mean	Std Deviations
FEV_1 % Δ	9.18%	22.2%	1.3
FVC % Δ	5.86%	11.98%	0.6
D_LCO (Hb corrected)	14.78	18.16	0.5
D_LCO % of predicted (Hb corrected)	55.81%	65.37%	0.4
V_A % of predicted	89.98%	98.14%	0.4

Table 4-12: Some of the attributes indicated as most related to the *FEV_1 Reversibility* class, for the 150 cases with *Obstruction*

There were 108 cases in the dataset with both *FVC Reversibility* and *Obstruction*. The class comparison between those cases and cases with *Obstruction* are shown in Table 4-13. Two of the same classes were identified as with *FEV_1 Reversibility*, although each to a lesser extent. Both were supported when looking at the attribute correlations (summarised in Table 4-14): cases with *FVC Reversibility* showed an association with RV, which increased from an expected mean of 130.51% of predicted to 150.57%. Cases with *FEV_1 Reversibility* showed no appreciable change in expected RV. In examining diffusion, cases with *FVC Reversibility* showed a small reduction in mean uncorrected D_LCO, dropping from 72.31% of predicted to 63.15%, although with a much smaller drop in corrected D_LCO (55.81% to 51.18%). This is the reverse of subjects with *FEV_1 Reversibility* which showed an increase in both those measurements. The differences are inconclusive, being no larger than half a standard deviation in either case, yet present an interesting result. Further analysis showed that for obstructed cases, the percentage of FEV_1 change bears no significant correlation to the D_LCO (expressed as a percentage of the predicted value); whereas the percentage change of FVC showed some association, with a stronger correlation for cases that have *FVC Reversibility* without *FEV_1 Reversibility* (correlation = -0.32474, $p < 0.05$).

Class	Cases	*p-sgain*
Evidence of Gas Trapping	27 (25%)	14 (52.9%)
Low D_LCO	53 (49.1%)	19 (36.4%)

Table 4-13: Classes showing the strongest association to cases with *FVC Reversibility*, for the 108 cases with *Obstruction*

Attribute	Expected Mean	Actual Mean	Std Deviations
FVC % Δ	5.86%	21.1%	1.6
FEV_1 % Δ	9.18%	17.69%	0.9
FEF_{25-75}% Pred Post-BD	44.66%	27.11%	0.7
FEF_{50} Post-BD	1.9	1.08	0.7
RV % of predicted	130.51%	150.57%	0.4
D_LCO % of predicted (uncorrected)	72.31%	63.15%	0.3
D_LCO % of predicted (Hb corrected)	55.81%	51.18%	0.2

Table 4-14: Some of the attributes indicated as most related to the *FVC Reversibility* class, for the 108 cases with *Obstruction*

A summary of several mean attribute comparisons between cases in the classes *FEV_1 Reversibility* and *FVC Reversibility* and are presented in Table 4-15. The results support previous indications that cases with *FVC Reversibility* have generally lower values for spirometry tests than cases with *FEV_1 Reversibility*. Interestingly the mean diffusing capacity (D_LCO) is worse in *FVC Reversible* patients than *FEV_1 Reversible*, even though the correlation between *FEV_1 Reversibility* and D_LCO was stronger than *FVC Reversibility* and D_LCO.

Attribute	FEV_1 Reversible	FVC Reversible
FEV_1% Pred Pre-BD	63.35%	57.46%
FEV_1% Pred Post-BD	75.64%	65.59%
FVC% Pred Pre-BD	82.49%	72.11%
FEV_1 / FVC Pre-BD	0.62	0.59
FEV_1 / FVC Post-BD	0.68	0.57
FEF_{25-75}% Pred Post-BD	61.84%	46.87%
FEF_{25} Pre-BD	3.56	2.29
FEF_{25} Post-BD	4.59	2.66
FEF_{75} Pre-BD	0.47	0.31
FEF_{75} Post-BD	0.65	0.31
FEV_3 Post-BD	3.01	2.2
SVC Post-BD	1.3	0.22
V_A / TLC	0.84	0.78
D_LCO% Predicted	67.72%	53.54%

Table 4-15: Significant differences between attribute means for cases with *FEV_1 Reversibility* and cases with *FVC Reversibility*

These identified associations and calculated results show the potential of the method to identify new or unexpected relationships for the data being examined, beyond what is found in a typical analysis. The analysis performed in the Agahi study was reproduced quickly with a larger dataset, and further relationships automatically identified to expand on the conclusions reached and the knowledge gained. An examination of the strength and value of the findings is presented in section 4.3.2.

Clinical Question 2: Can V_A be used to estimate TLC in patients with airflow obstruction?

Alveolar volume (V_A) is a closely related measure to Total Lung Capacity (TLC), such that it is often used to estimate the TLC. V_A is measured by the inhalation and holding of a known concentration of gas, typically helium, for a specific time limit; the amount of that gas that is exhaled is recorded, and the difference is recorded as the volume that the alveoli, and hence lungs, can hold (van der Lee, van Es, Noordmans, van den Bosch, & Zanen, 2006). TLC however is measured accurately by more complex means, such as a body plethysmography, in which the patient is typically enclosed in a sealed box of known air pressure; however the cost and size of the equipment make it a difficult and expensive test (Wanger, et al., 2005). Due to the differences in process, V_A underestimates TLC in patients with airflow obstruction, where the obstructive defect means that the measured gas cannot reach all parts of the lung (Ferris, 1978). However, if the extent of this effect could be estimated based on the degree of obstruction, or other factors, the less expensive V_A test could be used to estimate TLC effectively for all patients.

2007 Study Results

In Agahi's study, patients were grouped in 10% intervals of the percentage of the predicted FEV_1/FVC value, and the mean V_A /TLC were calculated for each. It was reported that an FEV_1/FVC ratio above 70% of predicted had an average V_A /TLC ratio close to 1; but below 70% the V_A /TLC ratio dropped progressively.

Further analysis was performed with patients exhibiting an FEV_1/FVC ratio < 0.7, and a regression equation defined for patients with an FEV_1/FVC ratio below 0.7, with R^2=0.252:

$$\left(\frac{V_A}{TLC}\right) = 0.53 + 0.49\left(\frac{FEV_1}{FVC}\right)$$

Data Analysis

An overall examination of the dataset showed a mean V_A/TLC ratio of 0.86, suggesting that V_A underestimates TLC in general. However, this may have been affected by an unusual number of abnormal patients, given the nature of the Austin Health dataset. To overcome this, only patients with *Normal Ventilatory Function* or *Normal Lung Function* were selected, which gave a mean of 0.91.

An initial comparison of cases displaying *Obstruction* to those without *Obstruction* immediately showed an association, with the mean V_A /TLC ratio dropping by 1.5 standard deviations to 0.78. Table 4-16 shows the findings of the comparison. For the data in this study, the information gain measurement identified 0.81 as the optimal cut point for predicting *Obstruction*, and 0.82 as the optimal minimum for predicting *Normal Lung Function* or *Normal Respiratory Function*. Similarly, the information gain statistic found that by selecting only cases with an V_A/TLC greater than 0.81 provided 80% more non-*Obstruction* cases than would be expected. This highlights that non-obstructed cases are much more likely to have an V_A /TLC ratio closer to 1, whereas the lesser improvement in cases with *Obstruction* suggests that while there is an association between V_A/TLC and *Obstruction*, it is not as reliable a correlation.

	Mean	**Optimal information gain**	**Gain improvement**
Normal Function	0.91	≥ 0.82	50%
Obstruction	0.78	≤ 0.81	59%
without *Obstruction*	0.89	≥ 0.82	80%

Table 4-16: Mean V_A/TLC, optimal cut point and improvement of that cut point for predicting the class from V_A/TLC

The calculated details for V_A and TLC showed no significant change in either mean for cases with *Obstruction*; but cases with a V_A < 93% of their predicted V_A showed a 42% information gain. TLC showed a 48% information gain for cases above 117% of predicted, but this was not supported by confidence or *p-sgain* measures. This indicates a relationship between *Obstruction* and a reduced V_A, but with no associated reduction in TLC. An investigation of cases without *Obstruction* showed a complimentary result: a 77% information gain increase for predicting non-*Obstruction* in cases with a V_A > 93% of predicted. The calculations also displayed a minor inverse effect for TLC: cases without *Obstruction* tended to show either a normal or reduced TLC value.

To investigate this relationship further, the cases were then divided into three subgroups based on their V_A/TLC ratios: V_A/TLC < 0.8; 0.8 < V_A/TLC < 1.2; and V_A/TLC > 1.2.

V_A /TLC < 0.8

Selecting cases with a V_A /TLC < 0.8 found 651 subjects (22% of the dataset). Immediately apparent was a very strong association with *Obstruction*, with 60.8% (396) of the set showing the class: 51.45% (203.75) more cases than expected for this subset. This is again a strong indication that cases with *Obstruction* have a low V_A/TLC ratio. A comparison for *Obstruction* representation between the different groups is provided in Table 4-17.

	Cases	***Obstruction***	***p-sgain***
V_A/TLC < **0.8**	651 (22%)	396 (60.8%)	+204 (+51.5%)
0.8 < V_A/TLC < **1.2**	2111 (71.2%)	452 (21.4%)	-
V_A/TLC > **1.2**	10 (0.4%)	2 (18.2%)	-
All cases	2963	875 (29.5%)	-

Table 4-17: Comparison between support, confidence and *p-sgain* values for different values of V_A/TLC and *Obstruction*

Overall, the measures indicated that cases with a reduced V_A /TLC show a general corresponding decrease in spirometric results, with a slightly weaker increase in lung volume results, and a weaker still decrease in gas transfer. Some of the stronger associations are presented in Table 4-18. The strongest correlation suggested was to the post bronchodilator FEV_1/FVC ratio. Of the two components, FEV_1 seemed to have a stronger association than FVC, although both seemed to be reduced. Post bronchodilator FEF_{25-75} showed a very similar reduction. The residual volume showed a marked increase in the subjects of this study. These factors together indicate a strong relationship between reduced airflow, such as is present in cases with *Obstruction*, and a reduction in V_A with no corresponding reduction in TLC; which is a logical result considering how the V_A is measured compared to how the TLC is measured, supporting the findings of Agahi's study.

	Expected Mean	Mean	Standard deviations	Optimal information gain	Information gain
FEV_1/FVC	0.74	0.57	1.3	< 0.54	73%
FEV_1 % pred.	84.24%	54.68%	1.2	< 64.5%	70%
FVC % pred.	94.1%	79.1%	0.9	< 82.25%	44%
$FEF_{25-75\%}$	2.79	1.3	1.1	< 1.11	60%
RV % pred.	108.7%	147.3%	0.9	> 137.8%	72%

Table 4-18: Attributes associated with the range $V_A/TLC < 0.8$

$0.8 < V_A/TLC < 1.2$

The next range examined contained 71.2% (2111 cases) of the dataset. No strong associations were found, except that 50% of the cases were listed as having *Normal Lung Volumes*.

$V_A/TLC > 1.2$

Only 10 cases (0.4% of the dataset) had an $V_A/TLC > 1.2$, with a maximum of 2.37. In general, these cases showed a slight increase in some spirometry, a general but ultimately insignificant increase in gas transfer, and a large mean drop in both TLC and RV which nevertheless showed no significant trend. It is expected that the minor statistical changes are products of the small sample size, and perhaps representative of outliers for TLC, or errors in TLC measurement.

The cases did show a very strong association with *Restriction*, correlating with 8 of the 10 cases. The two cases which did not show *Restriction* had V_A measurements much higher than predicted (230% and 130%) and reduced TLC (89% and 82% respectively). The only unexpected correlation displayed for those two cases was with BMI, displaying 32.04 and 36.39 respectively, both much higher than the average of 28.07; but with a sample size of two and no obvious pattern little can be drawn from this. A cursory examination of cases with a BMI above 30 showed no significant associations, nor did a combination of BMI and V_A/TLC measurement associate unexpectedly with any other class or measurement.

A quick examination of $V_A/TLC < 0.6$ showed a continuation of the trend shown by the three mentioned groups: *Obstruction* was further correlated with 71.1% of

cases (113 of the 159), this being 58.45% more than expected. A strong correlation with *Severe Obstruction* was also indicated; and by far the most correlated variable was the FEV_1 /FVC ratio, with a drop of 2.5 standard deviations from 0.74 expected to 0.41. Other spirometric measurements showed increased association, but none to the extent of the FEV_1 /FVC ratio.

Further Analysis

Based on these results it seemed evident that V_A provided a reasonable estimate of TLC except in the presence of *Obstruction*; and that a strong association was evident between V_A/TLC and FEV_1/FVC, with an increasing association the more disparate the FEV_1/FVC ratio. The V_A/TLC ratio was plotted against the FEV_1/FVC ratio, as displayed in Figure 4-6. As the figure shows, the data follows a linear model reasonably accurately with an R^2 value of 0.3583. However, as compared to the power trend line and the moving average, the linear model seems to overestimate for low and high values of FEV_1/FVC. The moving average in particular shows a fairly linear trend until approximately 0.7 FEV_1/FVC, at which point the trend flattens out to an FEV_1/FVC ratio of 1.

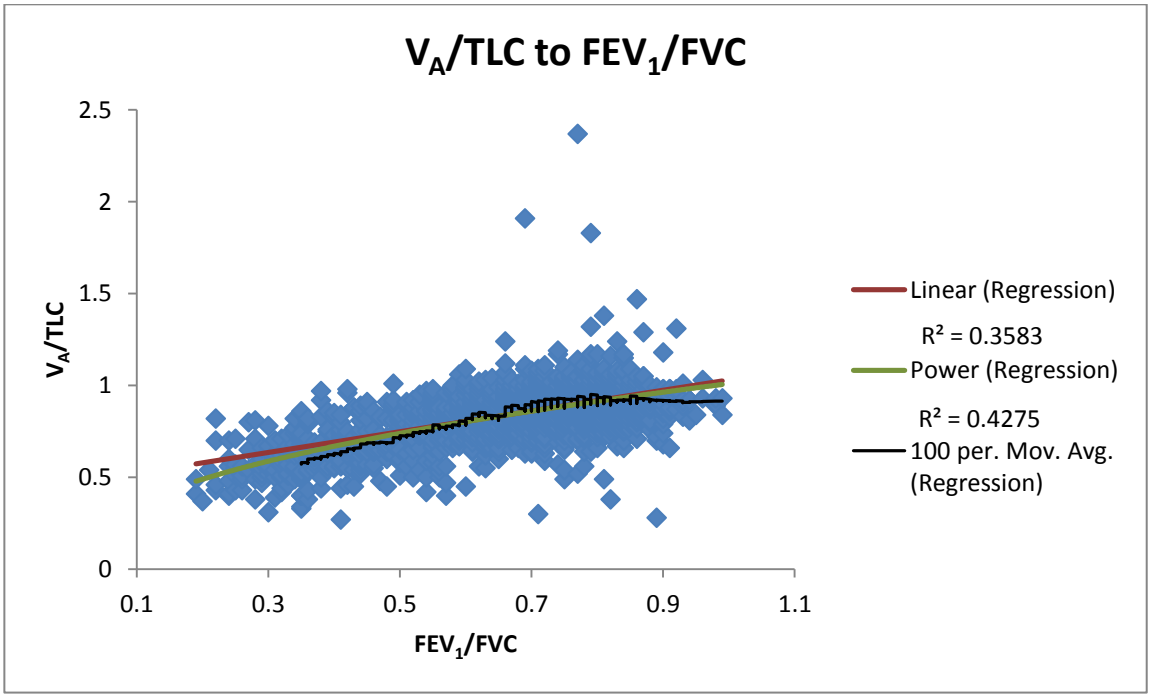


Figure 4-6: V_A/TLC plotted against FEV_1/FVC, showing a decrease in V_A/TLC of increasing magnitude as FEV_1/FVC decreases

From this evidence a regression analysis was performed attempting to predict TLC from a combination of the V_A and the FEV_1/FVC ratio. The data was divided around a threshold of a 0.7 FEV_1/FVC ratio, based on the indicated trend lines (and also as this equates to a common definition of *Obstruction*).

For the cases without *Obstruction*, based on pre-bronchodilator FEV_1/FVC ratio, generating a Pearson correlation coefficient between the V_A and TLC gave a value of 0.912; a generated linear regression model barely improved this to a value of 0.913 (R^2 values of 0.8315 and 0.8335 respectively), making any correction of V_A rather pointless. Post-bronchodilator FEV_1/FVC provided a correlation coefficient of 0.877, which a regression model again only improved by 0.001 ($R^2 = 0.772$).

The cases with *Obstruction* demonstrated, as expected, a much weaker correlation. Pre-bronchodilator values gave a correlation of 0.757, which a linear regression model improved to a correlation coefficient of 0.888 ($R^2 = 0.789$), with the equation:

$$TLC_{estimate} = -6.011\left(\frac{FEV_1}{FVC}\right) + 0.892(V_A) + 5.235$$

Post-bronchodilator values for obstructed cases provided a correlation coefficient of 0.787, which a regression equation improved to 0.901 (R^2= 0.812), with the equation:

$$TLC_{estimate} = -5.684\left(\frac{FEV_1}{FVC}\right) + 0.932(V_A) + 4.839$$

The post-bronchodilator regression equation provides a slightly stronger R^2 value based on less cases (n = 1873 post-bronchodilator, n = 2555 pre-bronchodilator), but as this requires a bronchodilator to have been administered to the subject, it may be less applicable in general circumstances.

To test the veracity of the system's selection of FEV_1/FVC as most correlated, a number of other indicated variables were used in place of the FEV_1/FVC ratio, including FEV_1, FEV_1 % of predicted and FEF_{25-75}; these were generally found to produce comparable, but ultimately less accurate, results.

Effects of BMI on Lung Function

As obesity is currently such a major health issue in the world today (Caballero, 2007), there is an increasing rate of studies trying to identify the effects of overweight and obesity on all aspects of lung function. Jones and Nzekwu

performed a study into the effects of BMI on lung volumes (Jones & Nzekwu, 2006), and Stritt and Garland studied the effects of obesity on volumes and spirometry (Stritt & Garland, 2009). Work in the area continues, such as with the study currently underway by O'Donnell et al on the effects of BMI on static lung volumes in patients with obstruction (O'Donnell, et al., 2011). These studies are by no means the only examples of such work.

Given these investigations into the relationships between BMI and lung function, this section will focus on using this system to reproduce the results of those studies mentioned, and on examining what other information the data may provide.

Previous Study Results

Jones and Nzekwu's study collected results for 373 patients both male and female with a range of BMIs, but with a number of other fixed criteria, including: over 18 years of age; an FEV_1/FVC ratio over 90% of predicted; a RV less than the upper limit of normal; D_LCO above the lower limit of normal, when adjusted for V_A. Results were analysed using linear or exponential regression, and analyses of variance (Jones & Nzekwu, 2006).

The study found linear relationships between BMI and VC, and between BMI and TLC, but without a significant change in either mean. FRC and ERV decreased exponentially as BMI increased, with the greatest rate of change in patients overweight or with mild obesity: at a BMI of 30, FRC was at 75% of the value of a person with a BMI of 20, and ERV at 47% (Jones & Nzekwu, 2006).

Stritt and Garland's study identified a lack of information regarding correlations between specific BMI levels and their effect on lung volumes, and even less information on correlations with spirometry. Patients were selected according to a series of criteria: a FEV_1/FVC ratio at least equal to predicted; a D_LCO at least 70% of predicted; and no evidence of respiratory muscle weakness. Patients were then grouped according to BMI (Stritt & Garland, 2009).

Reported results were that the 13 patients with a BMI below 30 displayed a mean TLC of 93% of predicted, a mean FEV_1 at 73% of predicted, and a mean FVC at 77%; whereas the 10 patients with a BMI at 30 or above showed 81%, 67%, and 70% respectively (Stritt & Garland, 2009).

O'Donnell et al have produced a larger scale study of 2,265 patients with a FEV_1/FVC ratio less than 0.7, investigating the effects of BMI on volumes and airflow function. They found that as BMI increases, FRC, RV, ERV and specific airway resistance decreased exponentially. They also found a linear decrease in TLC, and linear increases in IC and IC/TLC, but no significant effect with VC (O'Donnell, et al., 2011).

Testing Procedures

To perform a similar analysis to the Jones and Stritt studies, classes were defined in 5-step BMI intervals: *Underweight* (BMI < 20), *Normal Weight* (20 ≤ BMI < 25), *Overweight* (25 ≤ BMI < 30), *Obese I* (30 ≤ BMI < 35), *Obese II* (35 ≤ BMI < 40), *Obese III* (40 ≤ BMI < 45), and *Obese IV* (BMI ≥ 45). Before examining the data, the following criteria were added to remove extraneous factors: Age > 18; without *Obstruction* (FEV_1/FVC ratio ≥ 0.7); with an FEV_1/FVC ≥ 90% of predicted; normal D_LCO (above 80% of predicted); and RV < 120% of predicted.

New Results

No class associations were indicated for any of the BMI categories defined. Table 4-19 summarises the mean values for volume measurements, FEV_1 and FVC over each category (SVC showed no substantial or consistent change). Mean TLC showed a relatively consistent linear downward trend culminating in a 7.07% (0.7 standard deviations) drop in mean percent of predicted between normal BMI and a BMI above 40. FRC showed a strong decrease as BMI increased, dropping consistently until the last group. This trend also seemed to continue in the opposite direction, with the group of patients with a BMI below 20 showing a higher FRC. Using all BMI groups, comparing BMI to FRC (expressed as a percentage of predicted) gave a correlation coefficient of -0.38; although FRC appeared to improve slightly as BMI became very high indicating it may not be a linear trend (see Figure A-1 in Appendix A). ERV showed a similar relationship, including the *Underweight* group. Unfortunately a lack of ERV prediction equation in the data meant no percentage of predicted value was available. The ERV data also showed a slight improvement as BMI becomes very high, again possibly suggestive of a non-linear trend.

	TLC	FRC	ERV	RV	IC	FEV_1	FVC
Underweight	103.25%	106.77%	1.86	93.36%	108.4%	95.9%	98.3%
Normal Weight	104.33%	98.95%	1.64	93.05%	113.2%	100.4%	101.3%
Overweight	103.32%	91.62%	1.22	90.22%	118.4%	99.5%	99.7%
Obese I	100.85%	82.18%	0.82	90.74%	123.7%	96.1%	94.3%
Obese II	100.35%	80.24%	0.84	90.39%	128.1%	93.8%	92.6%
Obese III	101.25%	76.64%	0.64	92.9%	135.7%	94.5%	91.9%
Obese IV	97.26%	87.46%	1.04	92.31%	110.2%	94.1%	91.8%
BMI correlation	-0.146	-0.38	-0.5	-0.04	0.19	-0.16	-0.252
Significance	<.0001	<.0001	<.0001	0.27	<.0001	<.0001	<.0001

Table 4-19: Mean values for volume and spirometric measurements, for each of the defined BMI categories, expressed as percentages of the predicted value (no ERV predicted data was available, and so the direct measure was included)

The numbers show a small but insignificant decrease in RV as BMI increased, with subjects with a BMI between 25 and 40 showing the largest decrease (still only a drop from 93.05% of predicted to 90.22%). This was the only attribute of those examined that had no significant correlation, disagreeing with the findings of O'Donnell et al. IC showed a consistent increase as BMI increased, although subjects with a BMI above 45 displayed a sufficiently smaller mean. *Underweight* subjects also continued the trend with a decreased value. Although not displayed in the table, V_A/TLC showed a trend similar to other attributes, with a very small decrease from 0.95 to 0.9 between *Normal Weight* and *Obese III*, with *Obese IV* subjects showing an increase to 0.97.

These results correlated well with the Jones study, identifying similar relationships between TLC, FVC, and ERV. Notably the results here also identified the slight improvement for very high values of BMI, although the Jones study identified this improvement at slightly lower BMI levels.

In comparing the results in Table 4-19 to those of the Stritt study, there was no equivalent relationship found between BMI and TLC: Stritt found a drop from 93% of predicted to 81% when comparing patients without obesity to those with. The

divisions described here found a negative correlation, but not to the extent described by Stritt and Garland. Comparing normal (BMI between 20 and 24) subjects with those with a BMI above 45 found at most a drop from 104.3% to 97.3% (0.7 standard deviations). Comparing subjects with a BMI above 30 to those below, as in the Stritt study, again showed only a slight decrease as shown in Table 4-20. The general trend in this data appears to be a slight negative correlation between BMI and TLC, but nothing in the order of the data in Stritt's results.

Similarly, the system showed no significant change in the FEV_1, expressed as a percentage of predicted. Closer analysis showed a very similar trend of a slight decrease as BMI increased, with a small increase for very high BMI, as shown in Table 4-19. A direct comparison of BMI < 30 to BMI > 30 again showed a slight drop (0.36 standard deviations). Hence again, a trend seems to be evident, but not in the strength reported by Stritt. A very similar trend is apparent for FVC, although slightly more pronounced and with no increase as BMI becomes very large, although the rate of decrease slows significantly. The trend becomes smaller for post-bronchodilator FVC however (correlation coefficient -0.169). The percentage change of FVC also appeared to be correlated with BMI, increasing as BMI increases with a correlation coefficient of 0.206. FEF_{25-75}, FEV_1/FVC ratio, PEF, and FEV_1 post-bronchodilator change showed no correlation.

		TLC % pred.	FEV_1 % pred.	FVC % pred.
Stritt and Garland	**BMI < 30**	93%	73%	77%
	BMI ≥ 30	81%	67%	70%
This study	**BMI < 30**	103.7%	99.73%	100.24%
	BMI ≥ 30	100.67%	95.47%	93.7%

Table 4-20: A comparison of Stritt and Garland's results (Stritt & Garland, 2009) to those found from this data

Although the O'Donnell study is more comprehensive, the data available in this study still provided 875 more lung function reports to examine. Subjects were selected based on having the *Obstruction* class and a BMI greater or equal to 20, and were divided into the same BMI groups as previously. The correlations coefficients found in this study for the volumes measurements examined by

O'Donnell et al are presented in Table 4-21. Results are largely consistent with their findings, although IC showed no significant correlation, and SVC showed quite a weak correlation.

Attribute	Correlation	Confidence
TLC	-0.214	< 0.001
RV	0.189	< 0.001
FRC	-0.318	< 0.001
ERV	-0.301	< 0.001
SVC	-0.089	< 0.05
IC/TLC	0.206	< 0.001

Table 4-21: Correlation coefficients, with confidence values, for the lung volume attributes examined by O'Donnell et al (O'Donnell, et al., 2011)

4.3 Results and Discussion

In evaluating the success of the new method as a knowledge discovery tool for complex data, there are a number of considerations that bear an influence on the conclusions. These issues are discussed in the following section, before the evaluation itself is presented and conclusions made about the efficacy of the method.

4.3.1 Difficulties in Evaluation

The success of a knowledge discovery process is dependent on whether it discovers new and interesting knowledge, both of which are evaluated by a human analysing the results: the newness is dependent on the existing knowledge of the person performing the analysis, while interestingness is a subjective measurement that can depend not only on the knowledge and experience of the human but also on their insight, current thoughts, and contextual information (Clancey, 1993; Compton & Jansen, 1989; Liu, et al., 1997; Piatetsky-Shapiro, et al., 1994). This makes the effectiveness of a full knowledge discovery process inherently difficult to evaluate, dependent as it is on the human involved: while many studies have been performed comparing the effectiveness and efficiency of the data analysis component of knowledge discovery, where the rules produced (and reduced) can provide

quantifiable results (Freitas, 1999; Goebel & Gruenwald, 1999; McGarry, 2005), there is little consideration for evaluating the entire knowledge discovery process from beginning to end (Piatetsky-Shapiro, 2000; Pohle, 2003). These problems are particularly apparent in the approach described in this thesis, as the human involvement and guidance is an integral component at every stage.

This human involvement also causes the variability of results to be a stronger factor in this study than in many. Although major studies identified early that any knowledge discovery is an iterative process (Fayyad, et al., 1996b), research into data mining methods, and data-focused knowledge discovery methods, often base their results on a single pass of generation and results interpretation (Hidber, 1999; Lenca, et al., 2006; Marinica, et al., 2008; Tan & Kumar, 2001). This minimises the element of variability dependent on the person performing the analysis and result interpretation, allowing a simpler evaluation of results. In this method however, the strongly iterative approach and pervasive involvement of human expertise increase this variability.

This dependency on the human involved particularly complicates the evaluation of this study given the necessity of having a non-expert testing the system. The discovered knowledge is inevitably not of the complexity or quality that might be discovered by someone with experience and expertise working with the data; nor are the conclusions reached by interpreting the results as sophisticated. It is also likely that an expert would have much finer criteria for identifying interesting relationships, based on having more detailed expectations for what the data should represent. It is therefore difficult to evaluate the full capability of the method in identifying truly new knowledge. This implies that the method can only be evaluated here by testing what knowledge can be discovered that is new to the user; and, given other evidence, it might be extrapolated that someone with more significant expertise could derive newer and more complex knowledge. In some ways this makes evaluation easier: discovered knowledge that is new to a non-expert can be tested against existing literature, whereas truly new knowledge could not be otherwise verified. The significance of the discovered knowledge is discussed further in section 4.3.2.

Some work has been carried out attempting to determine how to best perform comparisons for knowledge discovery methods, but these methods invariably focus

on rule interestingness criteria and tend to ignore the human component. Works such as Freitas' 1999 study attempted to identify criteria that could be used to compare different rule identification methods, balancing efficiency with effectiveness; but while recognising the subjective aspect of interestingness, Freitas deferred the matter to other research (Freitas, 1999). Despite the work that has been undertaken in adding subjective interestingness measures to knowledge discovery approaches, there is still the significant question as to how to combine objective and subjective measures (McGarry, 2005); and until this is resolved, comparisons between methods which include differing levels of human involvement and subjectivity cannot be easily compared.

4.3.2 Evaluation of Approach

Given these difficulties, the method has been tested by using the system to resolve questions that were raised in the literature, or that were suggested by lung function experts as topics to be considered. This section will consider what can be concluded about the approach from the examination of those topics.

4.3.2.1 Discovered Knowledge

Each of the data analysis studies were performed in a single session, including analysing any relevant previous studies and interpreting the results; the system logs show a mean time of slightly less than 3 hours of use per study. Although more time could certainly have been spent analysing subsections of the data, qualifying results in regards to specific factors and finding further related studies, this form of exploratory analysis becomes an almost endless process with diminishing returns. Rather, the data analysis was continued until such a point as the user felt that new knowledge had been discovered and the research questions answered; which seemed a reasonable approach for any user of the system to take.

Examination of Results

This section will discuss the results of each of the data analysis studies in turn, qualifying the knowledge found with a subsequent review of relevant literature, to allow an evaluation of the method as a knowledge discovery tool.

Question 1: The distribution and pattern of lung function of subjects who met the FEV_1 and/or American Thoracic Society positive reversibility criteria.

The ratio of FEV_1 reversible subjects to FVC reversible subjects seems to be variable. Smith et al found 43 patients displaying FVC reversibility to 63 displaying FEV_1 reversibility (Smith, Irvin, & Cherniack, 1992), a ratio of 0.68; which exactly matches the ratio found in this study of 159 with FEV_1 reversibility to 235 FVC. However, this ratio was almost exactly reversed in a recent study from Saad et al (Ben Saad, Préfaut, Tabka, Zbidi, & Hayot, 2008), which found 49 FEV_1 reversible cases to 77 FVC reversible cases, a ratio of 1.57 (or 0.64 for FVC to FEV_1).

Smith et al compared the distribution of lung function for spirometry-derived reversibility to other means, but published no data on how different spirometry criteria compared. However, Saad et al published a comparison of FEV_1 and FVC reversibility to support their conclusion that FVC should more commonly be used to define reversibility. Their results showed a similar discrepancy between the two groups for mean FEV_1 % of predicted and FVC % of predicted. FEV_1 showed a mean of 46% of predicted for FEV_1 reversible cases, with 39% of predicted for the FVC group; a ratio of 0.85, comparing to a 0.87 ratio from this study. Saad et al showed a mean FVC of 69% of predicted for the FEV_1 group and 59% for the FVC group, a ratio of 0.86, consistent with this study's ratio of 0.87. The actual percentages of the predicted values are lower than those found here, but the ratios are internally consistent between the two studies. FEV_1/FVC showed no significant difference in the Saad study. SVC again showed a similar effect, with Saad reporting 70% of predicted for FEV_1 reversible cases to 62% of predicted for FVC reversible cases; this difference is of a much smaller magnitude than the one shown by this study, but nevertheless both show a pattern of reduced SVC for FVC reversible cases.

The claims of the Saad et al study that FVC is more sensitive in identifying reversibility (Ben Saad, et al., 2008) are not supported by the ratio of FEV_1 to FVC reversible cases found here; however, the number of cases which FVC identifies which FEV_1 does not, and the variety of cases and magnitude of effects in those cases, does suggest that FVC reversibility can provide important information about a case and should not be ignored.

Question 2: Can V_A be used to estimate TLC in patients with airflow obstruction?

A subsequent review of literature on this topic found a number of studies producing similar findings. Punjabi et al retrospectively analysed 2,477 patient results to assess the relationship between V_A and TLC. They also found that patients with a FEV_1/FVC ratio ≥ 0.7 showed a very strong correlation (V_A /TLC values between 0.97 and 0.99); whereas for patients with a reduced FEV_1/FVC, V_A generally underestimated TLC (V_A/TLC between 0.67 and 0.94). They produced a regression equation for a corrected V_A:

$$cV_A = V_A / \left[0.46 + 0.75\left(\frac{FEV_1}{FVC}\right)\right], if\ FEV_1/FVC < 0.7$$

(Punjabi, Shade, & Wise, 1998). Punjabi et al discuss a number of other studies publishing similar results: Burns and Scheinhorn (Burns & Scheinhorn, 1984) examined V_A and TLC comparisons in subjects with an FEV_1/FVC ranging from 0.28 to 0.95, and also found that an FEV_1/FVC ratio < 0.7 indicated a discrepancy between V_A and TLC. Similar findings for the relationship between reduced airflow and the difference between V_A and TLC have been shown by Ganse et al (van Ganse, Comhaire, & van der Straeten, 1970) and Ferris (Ferris, 1978).

Earlier studies have found differing results. Pecora et al found that VA produced an accurate assessment of TLC, and based on the derivation of regression equations and statistically significant correlation coefficients recommended that it be used in place of more expensive TLC tests (Pecora, Bernstein, & Feldman, 1968). Mitchell and Renzetti supported this result with their study (M. Mitchell & Renzetti Jr, 1968). Punjabi et al provide a lengthy discussion on the reasons for these discrepancies, justifying their conclusions (Punjabi, et al., 1998).

A more recent study analysing the relationship between alveolar volume and TLC concluded by defining an equation based on doubly correcting V_A using FEF_{25-75} and a measured difference between IC measured during SVC, and IC measured during the V_A measurement (Anees, Coyle, & Aldrich, 2009). This study retrospectively analysed 171 patient results, and concluded that their equation could be used to correct V_A for any patient recording a good effort in spirometry measurement. The results show the correlation between their doubly corrected V_A

and TLC with an equation for predicting doubly corrected V_A from TLC, with R^2 = 0.7145 (Anees, et al., 2009):

$$y = 0.9954x + 0.1047$$

These results are consistent with the analysis performed in this study, with FEF_{25-75} providing an effective correcting factor for V_A, although a comparison of the R^2 values might suggest, as found in the analysis here, that FEF_{25-75} is not quite as accurate a correcting factor as FEV_1/FVC. Qualifying the results by applying this study's FEV_1/FVC equation to the Anees et al 171 cases, or by applying the doubly correcting equation over this data should give a better insight, but the singular IC measurement in this dataset make this a future project.

Effects of BMI on Lung Function

The results of the analysis with BMI are more difficult to confirm as the literature inspiring the analysis is quite recent, and there have been no subsequently published studies. As was shown in the analysis however, the results generally matched the findings of the other published works, with some notable differences, and should provide valuable evidence supporting or extending their results.

4.3.2.2 Efficiency of Analysis

The efficiency of the statistics generation is an important consideration in ensuring that the interface is responsive to the expert's interactions, and that the expert can satisfactorily manipulate the data as they desire. The statistics used were kept simple to afford this freedom, such that the system can run through an online interface with a simple web server: even when considering the full dataset with a complex series of conditions the statistics page loaded in a few seconds. This delay was still noted to be an annoyance to users and a potential reason for not wanting to use the system for extended periods (this is described in section 5.4.2). There is however much room for optimisation in the generation of these statistics, in many areas including algorithmic optimisation, hardware upgrades, and a change of development platform or a shift to an offline interface. Increasing the size of the database or increasing the number and complexity of generated statistics and interestingness measures is still a very viable option. More complex interestingness measures can easily be adopted by only having them calculated on demand for a

section of data, as the system currently does with the information gain optimal range calculations. Such an approach does limit the level of suggestion and advice the system can offer, but the application and results of the information gain calculations used in the study show that such features can still provide useful assistance, especially in exploratory data analysis and the testing and validation of specific ideas.

4.3.2.3 Significance of Discovered Knowledge

It is important to be clear about the significance of the lung function conclusions that were discovered in this study, and those that might be found by using this approach. A search of recent respiratory literature shows that the discovered results of the different analyses are hardly groundbreaking in the field, but this is expected: the studies were performed by a non-expert in the field, with very little experience or knowledge of available literature and domain knowledge. The second question from the Agahi study provides an example: the results were certainly not ground breaking to the field, with a very similar result and regression equation having already been presented in a paper some 12 years previously. However, the results were new to the user, expanding the user's understanding of the data and the field. That the results were developed independently of similar conclusions, after only a few hours of analysis, indicate that the approach can be used to successfully discover new knowledge.

This is further supported when we consider that the user is relatively uneducated in the field, certainly having no formal education in the area and learning about the data only through the development of this study. It should also be noted here that any generalisation of the ease of use is somewhat lessened by the user also being the developer of the software, and hence being familiar with the interface and statistics used. However, the results shown by others using the statistical tools (presented in section 5.5.2) show few signs of difficulty. Regardless, the problem of usability is one that can be overcome with training and familiarity, perhaps combined with simple interface alterations. It is therefore expected that an analysis by a user with domain knowledge, research expertise, and specific questions and expectations in mind would produce far more interesting results.

Rather than discovering revolutionary knowledge for lung function, the primary goal was instead to demonstrate the intended application and the efficacy of the system. However, this does not mean that the results are irrelevant. The results still represent legitimate analysis of real lung function data, and as such they can provide useful evidence supporting or expanding on the results of other lung function studies, or provide directions for future research. The full significance to the lung function field of the specific results found here are not for this author to say; however, at the least they can be considered to show support for many of the findings of the cited studies, such as Saad et al (Ben Saad, et al., 2008) and Punjabi et al (Punjabi, et al., 1998); or to provide further evidence to develop the research of other studies, such as that performed by Anees et al (Anees, et al., 2009). The results of the BMI analysis may prove useful given the scarcity of currently published results (relative to the interest in the area), and the small number of subjects used in many of the relevant studies. However, there are a number of issues to be considered in generalising to prospective results that might be found by this approach.

A major factor in the applicability of the results found here is that they were developed using the percentages of predicted values. The use of fixed value interpretations is a major flaw, as the international standard is now to use statistically derived, individually calculated normal limits. This is a known problem with the system, which was unfortunately identified too late in development. Any further development or redevelopment of the method would incorporate statistically derived limits of normal from the beginning. However, this flaw does not detract from the efficacy of the system as a demonstration of the applicability of the approach. If incorporated from the initial stages of development the use of limits of normal would be a trivial change, not causing any change in the computation time of the system, nor in the applicability of the approach. Neither are the two methods (statistical limits of normal and percentage of predicted limits) mutually exclusive: both could be incorporated into one system concurrently. It is suggested that any change in the function of the system caused by the use of limits of normal would be only beneficial: finding more accurate results for lung function, broadening the range of experts that could happily participate, and allowing further comparative analyses of approaches to lung function interpretation, without a cost in

computation time of any significance. The use of statistically derived limits of normal should also reduce the complexity of the task, as experts are not required to remember or calculate relevant percentage limits.

The nature of the analysis as a retrospective study of archived data presents some concerns. This form of study is common in the health field, where collection of new data can often be difficult: studies such as Agahi's (Agahi, 2007), Stritt and Garland's (Stritt & Garland, 2009) and O'Donnell et al (O'Donnell, et al., 2011) are examples of such retrospective lung function studies, using previously collected or archived data to test ideas. As discussed by these studies, and has been discussed previously in section 2.5.1, the process of collecting the data for analysis can be difficult and needlessly complex. The approach developed here shows the potential for a unified database and a central knowledge base from which to perform such studies. However, this database as it currently exists does not conform to the more rigorous experimental designs of many of these studies: for example it does not have detailed information about the laboratories collecting the data or the equipment or processes they used, such as is present in most detailed lung function studies, for example Jones and Nzekwu's study (Jones & Nzekwu, 2006); nor does it contain data for all the attributes that might be of use, for example many cases lack D_LCO values corrected for haemoglobin, and many lack detailed smoking history, a critical factor for many areas of study.

Increasing the number of records in the database would also be a beneficial step. The numbers present in the current database are quite adequate for many studies, and should be enough to provide reasonable distributions for the studies described here. For studying rare cases however, a larger database or a specialist dataset would be required; for example, in the current database there are only 2 patients with a BMI above 30 and *Restriction*.

All of this extra information can certainly be added, and represents not so much a limitation of the approach, but a potentially limiting factor that must be considered in evaluating the significance of these results, and in making conclusions about the applicability of the method.

A further issue to be considered is that applicability of the statistics used to this type of data, and how well the results of the interestingness measures and calculated

information can be used to identify relationships. Most statistical calculations will not be perfectly suited to this data, due to the low degree of independence between each of the attributes. This is obviously well understood and accounted for within lung function and other health research, but it should nevertheless be noted.

The measures used in this system are only an illustration of the approach and demonstrate a fraction of the potential analysis power. If this approach were to be taken further the functions provided could be greatly expanded upon, to include more specialised analysis tools such as regression calculation, more extensive data mining calculations, data visualisations, and any other data analysis approach deemed useful for the domain of application. Each of these functions could add extra support for the user in data analysis, depending on what the user is trying to find or what may be present in the data. The computational complexity of the incorporated analysis methods must be considered, as intensive calculations will reduce the degree of interactivity; however, as with the optimal range information gain calculations used in this study, reasonably complex calculations can still be incorporated, with expert guidance selecting when they should be used.

As with all retrospective studies, this approach is hindered by the lack of flexibility of the database. As the data is necessarily de-identified and potentially from some years before the date of analysis, there are many specific questions that cannot be answered through this system due to a lack of the necessary data. This is a larger problem given that new forms of data will always be identified, new tests or procedures developed, and problems found that invalidate old data. The usefulness of a system in that design however is still evident from the many retrospective studies that are performed in health areas and the beneficial results that are found.

Given the stated limitations, the most appropriate use of the current system is as a source of preliminary data analysis: testing a hypothesis against the store of data and knowledge to verify that a trend is apparent. The current system is not best suited to a complete validation of hypotheses, nor for providing a conclusive explanation of any trend that it finds; but it can provide evidence and a basis for continuing research into a hypothesis, and preliminary suggestions of cause. The BMI studies provide examples of this sort of use. The other useful application of the current system is to find supporting evidence for existing studies that require additional data. A good example of this is Stritt and Garland's study (Stritt &

Garland, 2009), which presented interesting findings but based on small numbers of subjects (comparing a group of 13 patients to a group of 10). By testing their hypotheses and findings with this larger database, stronger supporting evidence was found to reinforce their findings and validate them against a larger database, which likely provided a better distribution. This provided qualifications for the exact numbers found, suggesting that a further and larger study is probably required.

4.3.2.4 Knowledge Acquisition

A significant feature of the new approach is the integration of a knowledge acquisition method, which means that the results of the data analysis, and any information generated from the process of that analysis, all feed into the existing knowledge base. This adds to the store of data available for analysis, improving the effectiveness of both the knowledge discovery process and the detail of the expert classification system. This is shown by the results of the data analysis studies: the definition of classes such as *FEV_1 reversibility* and *FVC reversibility* add this information to the cases which would not have been identified in any other way, yet which can provide significant information on the relationships between attributes and their meaning – in this case, that a capacity to reverse FEV_1 has a stronger correlation to a reduced diffusing capacity than FVC reversibility, among other correlations. This additional information is also retained, such that if any future data analysis study defines any set which has a significant relationship to these classes, this will be displayed to the user, thus adding to the information discovered about the new study.

The benefits to the classification system are more variable. The impact might be enormous in adding the identification of a newly discovered class of health problem or patient, providing a level of expertise to the system that some experts may not have. The benefits may be more subtle, such as providing a finer distinction between types of reversibility, which may influence a practitioner's decision on how to diagnose or treat a patient. However, as can be seen from the examples presented, if not checked the output would swiftly become cluttered with the subclasses and groupings specific to each particular data analysis effort: for example, an *Obese I with normal airflow, RV, and diffusing capacity* classification is likely to be mostly unnecessary information, and is presented in an unnecessarily

complicated manner. The use of separate knowledge bases and the ability to clone and transfer knowledge between knowledge bases restricts this pollution and still allows for the benefits, although it does require some administrative process to decide how the knowledge is allocated.

One of the main difficulties with analysing data such as that used in this study is the amount of prior knowledge required. It became evident in development that the volume of knowledge required for an effective data mining tool to be developed is far beyond what could reasonably be included by ordinary means. In identifying interesting relationships an option was implemented to exclude any attributes whose base attribute (for example FEV_1 in the case of FEV_1 % predicted, FEV_1 post-bronchodilator change, or FEV_1/FVC) were used in rule conditions. However, this immediately led to problems as genuinely interesting results were excluded: sometimes the relationship between FEV_1 and FEV_1 % of predicted, or FEV_1 pre- and post-bronchodilator, are exactly what need to be examined.

It was apparent that even using a method such as Liu's *general impressions* could not provide a reasonable solution for the general case. If using Liu's expectation-based measures, every attempt to use the system to answer a data analysis question would require a specialised knowledge acquisition process to identify the user's existing knowledge and expectations for the relevant segment of data in the particular context under consideration. As has been previously discussed, in the knowledge acquisition section of this study and in other studies, knowledge that is acquired can only be considered correct for the context it was acquired in; and even then it is subject to change (Compton & Jansen, 1989; Compton, et al., 2006; Richards, 2001). Attempting to acquire a knowledge base which can describe, for a complex domain such as lung function, the expectations of an expert for all attributes and for all contexts in which those attributes might be considered would be a considerable research task in itself. Such a knowledge base would likely describe the domain better than the knowledge base developed in this study and would doubtless prove a valuable resource for many tasks, but it is expected that it would also require a considerable commitment from a number of domain experts; a commitment that is beyond the scope and capabilities of this project.

The impracticality of predetermining detailed expectations is supported by the analysis performed here. For example, in identifying what relationships increasing

BMI has to lung function, the user has a specific set of expectations about what the data will show in the context defined in the analysis: subjects with normal airflow, normal diffusing capacity, a not unusually high residual volume, and an increasing BMI. Defining expectations for each individual attribute, for each possible context of this sort would be a much more time consuming process than simply entering those criteria and examining the data. The establishment of an exhaustive expectation knowledge base might be expected to allow more extensive data mining that may find relationships that the expert would not think to look for. However, the user's knowledge about relationships that they would not think to look for represent tacitly held knowledge, as do their expectations for those relationships. Acquiring such tacit expectations would require a considerable effort in knowledge acquisition, which is unlikely to reach completion and would likely still result in a large number of false positive results. In contrast, the method presented here allows the user to specify a context and have their expectations tested directly against the evidence (an important element in identifying tacit knowledge), with the benefit of having otherwise interesting relationships identified automatically. The efficacy is further enhanced by being able to adjust the interestingness thresholds for the context currently being examined, without being restricted to a set threshold for all contexts or having to predict the threshold that will best suit the current run of analysis.

The necessity for incorporating domain knowledge presents the biggest challenge to data mining for this data. Without being able to incorporate the level of knowledge used in performing the analysis here, the results of a data mining approach would be enormous (Liu, et al., 1997; Piatetsky-Shapiro & Matheus, 1994; Silberschatz & Tuzhilin, 1996). Examining the relationships that were discovered during the testing of the method, it was not immediately apparent from the defined interestingness measures that this relationship was significant. For example, in order to conclude that relationship of increased BMI and decreased diffusion was significant depended not only on identifying that a trend was apparent, but that one was not expected. In examining the various BMI classes for significant trends, an average of 43 attributes per class were identified by the system; of these, approximately a quarter were chosen as interesting and examined further, based on a tacit understanding of expectations and overlap between what each attribute represents. Were the system to automatically perform further analysis for all of these attributes,

this would lead to a much larger number of uninteresting and redundant results to be analysed.

Furthermore, even given the identification of such trends, nothing conclusive can be stated from them unless other potential factors are removed. The identification of these factors is a complex process dependent on the knowledge of the user: for example, in order to be able to say with any conviction that BMI has an effect on airflow, other potentially influencing factors need to be removed, such as whether the patient has COPD, asthma, or a similar problem. This in turn requires an understanding of which attributes indicate those problems, and what patterns in the data would represent patients that do not display those traits. The author is not aware of any knowledge discovery techniques that can sufficiently account for this issue.

Thus it can be seen that the necessity for complex domain knowledge in effective data mining is a major difficulty for data of this kind. The solution presented here is twofold. Firstly, to incorporate a knowledge acquisition process so the expert can define their expected results as a class; the system can then show the relationship between that class and the set resultant from the data analysis. Secondly, this approach does not overwhelm the expert with results that may or may not be significant, and instead allows them greater control over what data analysis is performed. This reduces the complexity of knowledge acquisition by allowing their knowledge to be applied more directly, which may improve efficiency given the well-recognised cost of knowledge acquisition (B. G. Buchanan, et al., 1983; Lenat, et al., 1985).

4.4 Conclusions

The results presented here show that this method can successfully perform a knowledge discovery task in a complex field such as lung function. While the lung function results derived in this study may not present anything likely to surprise a lung function specialist, they do demonstrate the efficacy of the system at allowing a user to discover new knowledge and develop their understanding of the field. That the results were confirmed by recent literature, and can produce findings relevant to current work, shows the capacity of the approach to derive useful knowledge. That the analysis of this data was performed and the subsequent results developed by a

relatively uneducated user, in a very short time frame, also indicates that the approach has a reasonable level of efficiency and simplicity. The analysis for each of the topics was performed without any specialised preparation of subjects or a clinical study, and with very little individual preparation.

It is not conclusively shown that this approach outperforms more traditional forms of knowledge discovery for this type of data; although this is suggested to some extent by the lack of published knowledge discovery works with this data. In particular, the complexity of the existing knowledge that was required to achieve the results found here appears to be beyond what other methods of knowledge discovery can effectively incorporate. As such, it is expected (though not proven) that it is the level of knowledge that can be integrated into the data analysis that allows the effective analysis of complex data. A further benefit is that any new knowledge can be immediately included in the knowledge base that this method is built around, by simply giving the current rule set a classification, and validating the knowledge against cornerstone case conflicts. The search conditions that were used to establish the case set of interest are converted into a rule and validated, via the normal MCRDR procedure. Once accepted, the rule is added to the knowledge base, and the knowledge contained therein will now be automatically applied to every case examined by the system. The new class, or expanded definition of an existing class, can then immediately be used in future analysis. Perhaps most significantly, this approach integrates the knowledge acquisition and results analysis components of knowledge discovery, allowing a smoother overall process of knowledge discovery. These tasks are often neglected in the development of knowledge discovery methods, despite evidence that they are costly components and vital to the success of discovering new, useful, and applicable knowledge, with many methods making no provisions for incorporating them (Fayyad, et al., 1996b; Kotsifakos, et al., 2008; Liu, et al., 1997; Piatetsky-Shapiro, 2000; Pohle, 2003).

Concerns with the experimental rigorousness of the data collection, and missing data elements, can restrict the conclusiveness of results found. This restriction, combined with the incorporated knowledge acquisition elements and the speed of the process, support the use of this system as an initial hypothesis validation or an exploratory data analysis tool. The results show that given a suggestion of a relationship or trend, a user can quickly use this tool to explore how well that

suggestion is represented in a dataset, and to then expand on their idea, exploring what related trends exist that might support, refine, or explain their hypothesis.

While the results indicate that the method can be used for knowledge discovery, the study used a combination of only two of the many dozens of datasets that exist, in Australia alone. A larger compiled set of data, with a more complete range of attributes, can only serve to benefit the efficacy of data exploration, the conclusiveness of results, and the range of applicability.

Similarly, while the data analysis functions incorporated into the system in this study demonstrate the potential for online analysis in such a format, these can be readily expanded upon to improve the effectiveness of the analysis and the level of assistance which the system can provide. The author also sees no reason why the knowledge acquisition components and the approach to exploratory analysis could not be incorporated into existing data analysis software.

Chapter 5 Knowledge Comparisons and a Tool for Learning and Assessment

5.1 Introduction

When acquiring and consolidating the knowledge of multiple experts in a given domain, there is potential for conflicts in knowledge to arise. In order to effectively consolidate the knowledge of all experts involved, these conflicts must be identified and resolved to each expert's satisfaction. While methods exist to collaboratively develop (Richards, 2009; Vazey & Richards, 2006) or integrate (Beydoun, et al., 2005) MCRDR knowledge bases, these methods do not focus on using evidence to resolve conflicts or, more importantly, on improving individuals' knowledge. This chapter presents a method to identify any conflicts, quantifiably measure the significance of each one with evidence, and present the reasons behind each conflict, such that the experts can reach a resolution and learn from the experience. This same pattern of quantified knowledge comparison is also applied as a novice learning and assessment tool, comparing a knowledgeable expert's input to that of a less knowledgeable professional or student. In addition to the benefits of knowledge comparison and the identification of weaknesses, it is shown that the knowledge acquisition process provides a useful opportunity for participants to apply learned theory and develop knowledge through practice. The contributions of this chapter, and their position in the larger method, are highlighted in Figure 5-1.

The learning outcomes of this approach are supported by the constructivist view of learning, which suggests that learning is an active process on the part of the learner: knowledge is not something that can simply be given or imparted, but needs to be developed based on an individual's interpretation and processing of experiences (Anderson, 2004; Duffy & Cunningham, 1996; Mezirow, 1991; Tapscott, 1998).

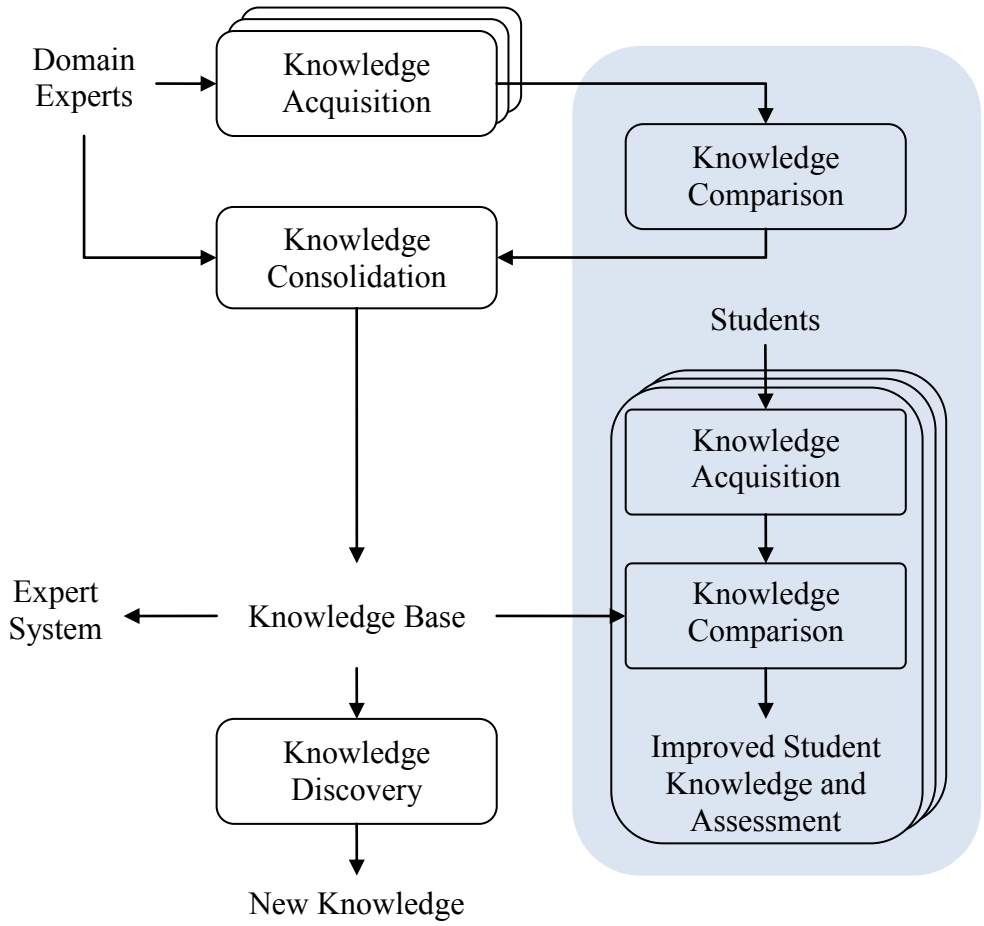


Figure 5-1: The methods presented in this thesis; the highlighted section shows the components presented in Chapter 5

5.2 The Learning Process and Constructivism

There are many variant schools of thought on the learning process. Behaviourism describes learning as an internal and unobservable process, asserting that learning can only be described by observing the learner's subsequent behaviours (Anderson, 2004; Good & Brophy, 1990). Cognitive psychologists attempted to describe learning, creating a model dependent on memory, motivation, processing and reflection (Anderson, 2004; Craik & Tulving, 2004). They asserted that there are 3 levels of memory, sensory memory, short-term memory, and long-term memory, and that the level of thought, or processing, given to knowledge will decide how it moves from one memory store to the next. Knowledge is then encoded in long-term memory in the form of networks of concepts, or information maps (Anderson, 2004; Stoyanova & Kommers, 2002).

A more recent school of thought is that of Constructivism, which takes a similar view to cognitive psychology but with a stronger focus on the experiences and perceptions of the learner. Constructivism states that knowledge is not something that can be given to a learner, but requires that the learner be much more active in the process: knowledge is only developed based on the learner's individual interpretation and processing of their experiences (Anderson, 2004; Duffy & Cunningham, 1996). While a learner may be told or given a concept, that concept will not form into knowledge unless the learner has an opportunity to apply it to a particular context or example, witnessing how the concept functions in a concrete, contextualised fashion. This gives rise to the goal of *situated learning*, whereby knowledge should be learned in the same context that the knowledge will be applied (Lave & Wenger, 1991). There is also an emphasis that knowledge is something which is discovered by the learner, rather than taught by a teacher (Tapscott, 1998). In support of this view, evidence and practical application have long been identified as beneficial in the development of understanding, as opposed to rote learning (Brown & Palincsar, 1989).

This suggests that learning should involve practical elements, as learners will better be able to form knowledge if presented with evidence to support theory and are allowed to discover the results of the application of theory over real examples (Duffy & Cunningham, 1996; Lave & Wenger, 1991).

Both the constructivist and situated cognition views of learning hold that knowledge is dependent on the context in which it is described (Anderson, 2004; Duffy & Cunningham, 1996; Mezirow, 1991; Tapscott, 1998). Previous studies have shown that different experts can present different results when asked the same question in different circumstances, even when having the same underlying beliefs, and that these conflicts of knowledge can equally occur from a single expert describing their knowledge in different ways, or from a difference in the underlying beliefs of two experts (Compton, 1992).

Based on these models of learning, it is clear that the practical application of knowledge is a critical component in effective learning, and that examples and evidence are integral to this process. In light of this, the method presented here makes use of evidence as much as possible when comparing knowledge and

assisting in the resolution of conflicts, with the goal of improving the knowledge of those involved.

5.3 Methodology

In Chapter 3, the expertise of multiple experts was acquired in separate knowledge bases, using the MCRDR knowledge acquisition framework described in that chapter. In order to effectively and accurately combine the acquired knowledge bases, including identifying and resolving conflicts, a strategy was implemented taking advantage of the large database of cases available.

Rather than only comparing the conceptual structures present in the knowledge base, the database of cases allows an evaluation of how close the two knowledge bases are in practice. The fundamental principle is to compare how the two knowledge bases function over a large set of cases, which should highlight the differences in definitions and provide a quantifiable measurement for how different each definition is. This focus on evidence is especially relevant considering the viewpoint that knowledge is only correct in the context that it is acquired for, and may change when discussed in a different context: in order to accurately compare definitions, and especially to resolve conflicts, evidence is required to provide sufficient context.

5.3.1 Knowledge Consolidation

5.3.1.1 Testing

The knowledge comparison method was first tested with the knowledge bases of the three experts described in Chapter 3, in order to develop a consolidated knowledge base for general use. As described in section 3.2.1, one lung function expert developed a knowledge base independently, and two others collaborated on a single knowledge base. Through a combination of some domain knowledge, identification of similar rule conditions, and consultation with the experts involved, the system administrator (the author) identified as many classification equivalencies as possible while attempting to preserve detail.

5.3.1.2 Equating Classifications

The first stage in the method is to identify equivalent classifications between the two knowledge bases. While this step could be avoided by finding an agreed upon standard of terminology before the knowledge acquisition process, this was not done in order to avoid limiting the level of knowledge that was acquired. It was also desired to keep the process as natural to the expert as possible, and to acquire the experts' terminology as much as possible, as a terminology comparison may itself provide interesting results.

A post-acquisition definition of equivalent classifications can be partially automated, by an analysis of the rule conditions used to reach each classification: any classifications that use the same rule conditions can be considered very likely to be synonymous. This automation could further be extended by considering very similar conditions, or by identifying classes that include the same set of cases. None of these options were implemented in this study, however, as there was no expectation that this domain would provide vastly different terminology, or that synonymous terminology would be difficult to identify. In such domains the grouping of classifications can be performed manually, through application of domain expertise, consultation with the experts involved, and examination of rule conditions.

5.3.1.3 Quantified Comparison

To generate a quantified comparison between the two knowledge bases, a simple algorithm is followed. Each case in the dataset is examined in turn, and the results for that case are compared between each knowledge base (either by performing an inference through the knowledge base with that case, or by recalling the stored results for that case). Various elements of the results are recorded: the relevant counts of occurrences of each classification (taking into account the defined equivalencies); the unique occurrences for each knowledge base; and the matching occurrences. This system also recorded the number of cases in each knowledge base having each quantity of classifications, for example the number of cases that had two classifications, the number of cases that had three classifications, and so on. This allows the calculation of the average number of classifications per case in each knowledge base, and the percentages of matches, unique occurrences, and conflicts,

both for each classification and for each knowledge base overall. Although more statistical measures could be derived from this process, and more measurements recorded, these were deemed sufficient for the purposes of the comparisons in this study.

5.3.1.4 Interface

In this system, the first statistics shown are the number and percentage of matching cases and conflicting cases, both with and without classification equivalencies, and the mean percentage of classifications matched per case. Next are the frequency of numbers of classification per case, and the mean classifications per case, for each knowledge base. Finally, each classification equivalency group is listed, and for each one the number of cases which have that class in each knowledge base, the number of matches, and the number of unique occurrences are shown. Each of these also has the relevant percentage that the number represents. Options are also presented to view the rules for each knowledge base which lead to those classifications.

An important component of this process is the ability to view the cases relevant to any particular comparison. Whenever a set of cases is described, the numbers displayed provide links to allow the user to view those cases. While the particular interface used is irrelevant, the function of viewing and allowing action on the described cases is integral to the usefulness of the comparison method.

5.3.1.5 Conflict Identification and Resolution

Once equivalencies have been defined and statistical comparisons made, the task is to identify the differences between the experts' rules, the causes of those differences, and how they might be resolved to reach a satisfactory consensus. The grouped classifications are sorted in order of significance of difference and are worked through in turn, by first examining the rules to see if the differences were caused by an error or a difference of opinion. Where necessary the experts are consulted to establish if an error has been made. If not, the experts are informed that there is a disagreement between them, and a discussion initiated to determine exactly why the disparity exists, which would be the best alternative to use, or what other options are available. This process is aided by the presentation of the statistics showing how

significant the differences are, how the different rule definitions relate to other definitions, and summaries of the attributes of the cases for each definition. Further assistance in resolution can be found by presenting the experts with exemplar cases which display the conflict, to ensure that the experts have a genuine conflict of opinion, and to ascertain exactly which attribute each expert uses to define that conflict. This process is summarised in Figure 5-2.

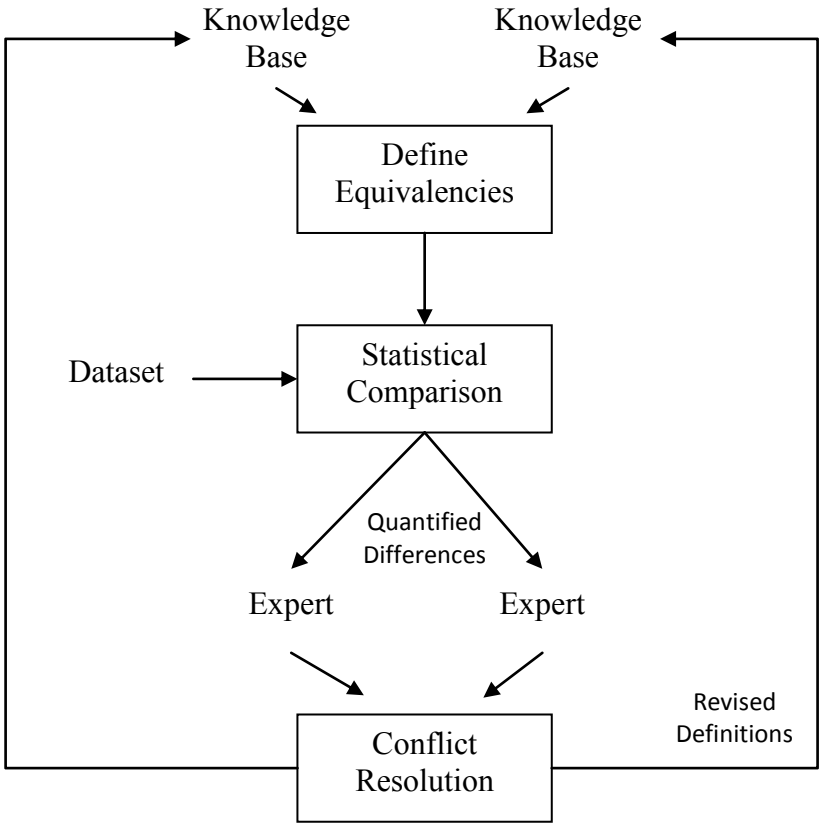


Figure 5-2: Summary of the conflict identification and resolution process

It should be noted that this is as iterative process: it is likely that after initial equivalencies are defined and comparisons examined, some equivalencies may be found to be incorrect or further equivalencies may be required. Some experts may wish to add further classifications, or change or remove previous classifications. This can be accommodated, equivalencies adjusted as necessary, and the statistics recalculated.

5.3.2 A Learning Tool

It was hypothesised that this same strategy could also be applied as a learning tool for domain novices, and potentially as an assessment tool for students, with only slight modifications. The system could be applied such that it provided students with a means of testing their lung function interpretation knowledge in a practical situation, with real lung function tests, and receive feedback about their comparisons to an expert level of knowledge. This feedback could identify the differences in classifications between the student and expert, and for equivalent classifications could identify the rule conditions that differentiate them. The system could again make use of real cases in demonstrating practical examples in which the student's definition disagrees with the experts'. Such a comparison might also effectively be used as an assessment tool in various ways: by determining the number of cases where the student's knowledge matched the experts', both for all cases and for only those seen by both; calculating the degree to which each case matched; and by examining how different their definitions are. The system could also provide feedback on which specific areas the student's knowledge is lacking.

5.3.2.1 Compact Knowledge Acquisition

Carrying out a full knowledge acquisition process with multiple experts, and many cases, would however be too time-consuming. In order to practically test the approach, it was decided that a limited set of cases would have to be defined. 20 cases were selected as being roughly representative of the spread of classifications non-specialists might be expected to reach, based on the experts' definitions and with expert consultation. As well as generic exemplars of classifications, some borderline and difficult cases were chosen to attempt to force participants into making precise definitions.

5.3.2.2 Testing

To test this application, a range of participants were sought. The first type of users tested consisted of medical students, ranging in experience from the third year up to the sixth (and final) year of the degree. These students were invited to participate via a group email to all medical students at the University of Tasmania from third year and higher. All willing respondents were included in the testing. Ethics

approval was obtained to interview participants, and consent forms signed prior to participation.

Each participant was informed of the goals of the study, how the study would be carried out, and what results might be found. They were then directed to and given a username and password for the online system, but wherever possible the participants were met in person and guided through the knowledge acquisition process, in order to observe the process, answer any questions, and to ensure everything ran smoothly.

Before beginning, each participant was given a questionnaire to ascertain their level of experience and confidence with interpreting lung function tests, including a sample lung function report to interpret on paper (see Appendix B). Upon logging in to the system, each participant was directed first to an online tutorial for how the system was structured, how to view cases, how to enter their interpretations (classifications) for each case, and how to define rules to justify their decisions. Participants were then directed through a MCRDR knowledge acquisition process, by examining each of the 20 sample cases in turn, for each one describing their classifications then justifying them by defining rules for each classification. Each participant developed their own knowledge base, independent of any other acquired knowledge. Where the knowledge acquisition involved a face-to-face meeting, the participants' actions and difficulties were discussed and observations made, with care taken to avoid interrupting the process as much as possible. It was made clear to the participants that the interviewer was only there to help with technical issues and resolve confusion about the process, and that the interviewer had no specialist knowledge in lung function and should not be taken as a guide for any knowledge-based decisions. It was also made clear that the participation was anonymous, and that it would not contribute in any way to their assessment. Once the 20 cases were completed, each participant was given another questionnaire asking their opinions on the software, the information presented, the usefulness of the process, and if the student felt they had learned anything (see Appendix B).

Once each participant had developed their knowledge base, each one was compared to the compiled expert knowledge base developed previously, via the same methods as described above. However, rather than attempt to resolve conflicts, an individual report was generated for each participant contrasting the participant's knowledge

with the experts', and provided to the participant if desired. The knowledge comparison was expected to be more difficult due to the range of difference in knowledge: the experts were expected to define many classifications which the students could not. While it was thought that the comparisons may be more effective if a knowledge base was developed representing a "perfect student", no such knowledge base was available. Also considering that the participants are at differing levels of education and hence expected to have differing levels of knowledge, the expert knowledge base was used instead. This was considered as possibly having benefits for learning however, as feedback can be more detailed and complete, closer to a practical knowledge scheme rather than a generalised reproduction of textbook patterns. To assist in ameliorating the discrepancy in knowledge, a component of the comparison process was the limiting of classifications to only those that appear in both knowledge bases, to allow a measurement of the correctness of the participants' knowledge.

5.4 Results

5.4.1 Expert Knowledge Consolidation

The comparison of the two expert-developed knowledge bases discussed in section 3.2 demonstrates the application of this method for knowledge comparisons. While the practical outcomes of this comparison were discussed in that chapter, the process of that comparison will be presented and discussed here.

5.4.1.1 Equating Classifications

Fundamentally, the process of equating classifications is a process of analysing the terminology. The terminology used by the experts in their knowledge bases will be briefly discussed to provide some context to the equating of the classifications, then the groupings themselves discussed.

Classification Terminology

It was found that there were significant differences in the terminology used for classifications in this study, not only between the three experts, but within the input of each expert individually. The variability was ameliorated to an extent by the document detailing the standard classifications expected to be used (see section

3.2.1), which was circulated among the experts involved, and provided a good reference base. However when this same expert later defined his own knowledge base, although he largely used the same rules, he used different terminology or different forms of terminology for almost every classification.

On investigation of the classifications used it is immediately apparent that there are, as with most language structures, many ways of expressing each classification. This stems in part from the nature of the classifications: they are not comprised of exact medical terms or final diagnoses, but are interpretations of the information available, meant to summarise what the body of test results represent and aid in a final diagnosis. Table 5-1 shows some examples of the differences in classifications, with classifications taken from all experts (/ indicates two distinct classifications).

	Classification(s)	**Alternative Classification(s)**
1	Obstruction – Mild	Mild airway obstruction
2	Low D_LCO	Impaired gas transfer
3	Mild Obstruction/Reversibility	Mild airflow obstruction with a positive response to BD
4	Hyperinflation	Moderate Hyperinflation/Severe Hyperinflation

Table 5-1: Terminology Differences

Table 5-1 displays three distinct problems with terminology that were encountered. The first three comparisons are examples of semantic differences: *Obstruction – Mild* and *Mild airway obstruction* are expressing exactly the same interpretation, and were at times used as the classifications for identical rules. Similarly, *Low D_LCO* and *Impaired gas transfer* express the same classification and are reached by identical rules: and both were defined by the same expert. The third comparison shows another variation on the first, but with some extra information added; this extra information is added by a distinct classification in the first instance, *Mild Obstruction* and *Reversibility*; and by a single classification in the second, *Mild airflow obstruction with a positive response to BD*. This terminology difference is slightly more problematic than the semantic differences shown by the previous examples. The last example shows another problem, but a common one: in the first instance an expert described *Hyperinflation* as a classification, whereas in the second the expert went to a further level of detail by defining *Moderate*

Hyperinflation and *Severe Hyperinflation*. It was noted that this problem frequently occurred between different experts, although both these instances were again created by the same expert.

Defining Equivalencies

Although the terminology can vary greatly between experts (for example, entries 2 and 3 in Table 5-1), it is largely irrelevant which versions are used. As mentioned, many of the more distinct differences did in fact come from the same expert, and at no point did any participant, in any part of this study, express or display confusion about the terms used. All of these terms seem to be equivalently understood by any level of expert in the field, being merely semantic variations on expressing the same underlying concept: provided that the expert has enough training and experience to understand that concept.

As described previously, given the relatively limited number of classifications in use (61 distinct classifications between the two knowledge bases) the classification equating was resolved by manually identifying which classifications were equivalent, and marking them as such. For purely semantic differences this was a simple process, although it involved consulting the experts occasionally. For compound classifications, the multiple classifications would be equated to the single other classification. When experts used different levels of detail, all associated classifications were equated to the most general classification.

In all, 30 of the 61 classifications were equated, into 5 different groups, with half of these in a single group. This group, all related to *Obstruction* classifications, was particularly problematic as it contained many gradations of severity, and an extra component that was sometimes added as a separate classification and sometimes compounded into the main classification. The severities presented difficulties because each expert had their own definitions of what the rule for each severity should be; and the independently developed knowledge base had a different number of gradations (6 compared with 3) that did not neatly overlap with the others .

The compound element was the presence of *Reversibility* or a *Positive response to BD* (bronchodilator) in the patient. This element of the classification is simple to define as its own classification, as the expert mostly did in the independently developed knowledge base. However, as the reversibility classification is most

frequently only considered in the presence of an obstructive classification, the natural tendency for experts is to define a compound classification with both elements, as in the form *Reversible mild obstruction*. Experts were also inconsistent with when they would include the reversibility element and when they would not, for example the same expert might define *Reversible mild obstruction*, *Irreversible mild obstruction*, and *Moderate obstruction*.

The second largest grouping of classifications, containing 6, involved those displayed in the second example of Table 5-1: *Low D_LCO* and *Impaired gas transfer*. In the independently developed knowledge base these equivalent classifications were graded into severities *Mildly/Moderately/Severely impaired gas transfer*, whereas the collaborative knowledge base used the *Low D_LCO* term, along with an exception rule qualifying the *Low D_LCO* classification in certain circumstances. Given the gradations were present only in one knowledge base they were equated for comparison purposes.

Two other groups, the *Restriction* and *Hyperinflation* classification groups, were again grouped because the independent knowledge base included gradations (2 and 3 extra grades respectively). Finally, in the collaborative knowledge base the classification *Normal Lung Volumes* had an exception, *Normal TLC*, qualifying the classification in circumstances where the expert determined it was not strictly correct. As no similar detail was included in the independent knowledge base, the two were equated for comparisons.

5.4.1.2 Comparing Results

Before Equating Classifications

When initially compared, because of the differences in terminology and levels of detail used between the experts, the collaborative and independent knowledge bases had no cases with exactly the same classifications. If only classifications that appeared in both knowledge bases were considered, 40% (1194) of the cases matched, with an average 12.3% of classifications matching per case; however with each knowledge base providing a mean 2.7 and 3.8 classifications per case, this gives an average of less than one matching classification per case, and in fact means that most cases have no matching classifications. Without grouping the

classifications, the collaborative knowledge base used 27 different classifications, 19 of which occurred only in that knowledge base. The independent knowledge base used 41 classifications, 33 of which were not used in the collaborative knowledge base.

After Equating Classifications

After defining the classification groupings, the collaborative knowledge base contained 21 distinct classifications or classification groups, of which 13 did not appear in the independent knowledge base. The independent knowledge base itself contained 23 classifications or groups, 15 of which did not appear in the collaborative knowledge base. Both these and the pre-equated classification numbers are presented in Table 5-2 for comparison.

	Independent		**Collaborative**	
	Class Groups	**Unique**	**Class Groups**	**Unique**
Before Equating	41	33	27	19
After Equating	23	15	21	13

Table 5-2: Total number of classifications or classifications groupings in each knowledge base, and number of classifications or groups that occur in only one knowledge base, before and after equating classifications

The comparison of the knowledge bases over the full dataset is summarised in Table 5-3. The comparison showed that there were still only 5 cases with perfect matches, with 99.8% having some difference in classifications due to the large number of unique classifications defined in each knowledge base. Some 36.9% (1091) of the cases were designated ―weak matches" (cases which match when only considering classifications that had an equivalent classification in both knowledge bases). The average number of classifications matched per case doubled to 24.6%, although still with an average of less than one matched classification per case.

	Matching Cases	"Weak Matches"	Average Classifications Matched per Case
Before Equating	0	1194 (40.3%)	0.4 (12.3%)
After Equating	5 (0.2%)	1091 (36.9%)	0.8 (24.6%)

Table 5-3: Cases with equivalent classifications and the mean number of classifications matched per case, between the collaborative and independent knowledge bases, before and after equating classifications

To aid in the comparison of the two knowledge bases, the frequency of number of classifications per case was calculated, and is displayed in Table 5-4. This suggests a number of points about the relative detail of the different knowledge bases. The independent knowledge base clearly tends to go to more detail for each case, with a mean average of one more classification per case, and a roughly equivalent modal average. The independent knowledge base might be said to be more complete, as it has far less cases with no classifications, and a strong majority (over 90%) with 3 or more classifications. It also has a much larger number of cases with high numbers of classifications, with almost 24% of cases having 5 classifications or more. These numbers may suggest however that it has defined some simpler, more general classifications that apply to large numbers of cases; a conclusion which is in fact supported by the individual classification statistics examined shortly.

	0	1	2	3	4	5	6	7	8	Mean per case
Collaborative	146	52	1297	859	491	118	0	0	0	2.6
Independent	5	95	193	1081	884	479	170	48	8	3.7

Table 5-4: Frequency of number of classifications per case, for each knowledge base

The most useful statistics generated are those about each particular classification grouping. After defining equivalencies, 8 classifications or groups of classifications were identified as appearing in both knowledge bases. The results of these comparisons are summarised in Table 5-5. Described are the total number of cases receiving each classification (~~o~~ccurrences"), the number of cases for which the classification appeared with both knowledge bases (~~m~~atches"), and the number of

cases for which the classification was applied in only one knowledge base (~~–u~~nique occurrences"). For example, 1991 cases were found to have the equated classifications *Normal lung volumes* or *Normal TLC*; of these, 84.8% (1689) had the classification in both knowledge bases. Some 227 cases had one of the classifications only in the collaborative knowledge base, and 75 cases had the classification group only in the independent knowledge base. This indicated both that the contrasting definitions for this classification group needed to be examined, and also the magnitude of the difference.

Classification Group	**Occurrences**		**Matches**	**Unique Occurrences**	
	Collaborative	***Independent***		***Collaborative***	***Independent***
Normal Lung Volumes	1916	1764	1689 (84.8%)	227	75
Impaired gas transfer	1287	847	842 (65.2%)	445	5
Obstruction	454	871	442 (50.1%)	12	429
Hyperinflation	256	568	256 (45.1%)	0	312
Restriction	569	298	245 (39.4%)	324	53
Evidence of gas trapping	888	136	136 (15.3%)	752	0
Small Airway Obstruction	94	171	41 (18.3%)	53	130
Normal TLC but evidence of functional hyperinflation	15	150	15 (10.0%)	0	135

Table 5-5: Comparison results for the classification groupings which appear in both knowledge bases

In support of the theory that the independent knowledge base includes some simpler and more general classifications than the collaborative knowledge base, the

comparison also showed a series of classifications unique to the independent expert's knowledge base which have high numbers of cases. The classification *No evidence of gas trapping* occurred only in the independent knowledge base, for 1929 cases (65% of the dataset). While the opposite classification *Evidence of gas trapping* appeared in both, it is at least partially this explicit definition of the negative form that inflates the numbers presented in Table 5-4. The definition of *Normal spirometry* (1216 cases) and *Normal gas transfer* (455 cases) are similar. Other definitions inflating these numbers were *This patient is underweight* (151 cases) and *This patient is obese* (808 cases), both quite general classifications used primarily to summarise data and help inform more complex decision making. Other classifications such as *Evidence of non-uniform ventilation* (823 cases), *Improvement in FVC post BD* and others showed some more complex classifications that are only reached in this knowledge base. There are also some classifications at a similar level of complexity that occurred only in the collaborative knowledge base. Discussions with the experts found these discrepancies to be representative of different specialisations or points of focus for each expert. The higher level of complexity for the independent knowledge base may also be a result of the lengthier knowledge acquisition process undertaken by that expert, viewing more cases and hence revealing more tacit knowledge.

5.4.1.3 Evidence-based Conflict Resolution

The most important aspect of these statistics here is their application in combining the two knowledge bases. As was described in section 3.3.1, the identification of these conflicts, and the presentation of the relevant statistics to the experts involved, facilitated directed and detailed discussions into where the experts disagreed, assisted by the indicated significance of those disagreements. The same methods used to identify those conflicts, and the exploratory tools described in section 4.2.2, allowed the impact of potential modifications to each expert's knowledge to be trialled and the results presented to find the best possible solution. Section 3.3.1 describes the five conflicts which did not have obvious solutions, but required discussions with the experts. Once presented with the conditions of their differing rules, three conflicts were resolved immediately. The remaining two differences remained in conflict until the statistics pertaining to those rules were presented. The relative impacts of the differing rules, defined by the number of classifications

generated by each, led to one conflict being resolved; the second was corrected after a more detailed examination of the impact of the differences, facilitated by the statistical methods described in section 4.2.2.

5.4.2 Novice to Expert Knowledge Comparisons

The results for the student knowledge acquisition and knowledge comparisons are presented here, for each student in turn. The implications of the results, and a discussion on the efficacy of the method as a learning and assessment tool, are presented afterwards in section 5.5.2.

5.4.2.1 Student 1

The first student participant was in their third year of a medical degree, and as such was expected to display a limited understanding of lung function. The student described that they had ~~–s~~ome" confidence with interpreting lung function reports (3 on the 5-level Likert scale), and estimated that they had seen less than 10 reports previously. Informally, the student professed from the outset to having little understanding of lung function reports, and that they were participating primarily to gain some experience with examining such reports. Some bugs and un-optimised code were evident during this knowledge acquisition session, slowing the process somewhat.

Terminology and Equating Classifications

After examining the 2 cases, the student defined six different classifications: *Restriction*, *Restriction with impaired gas exchange*, *Normal lung function*, *Obstruction – Mild*, *Obstruction – irreversible*, and *Obstruction – severe, irreversible*. The terminology itself shows a quite limited range of knowledge, being contained in a small number of classifications with repeated themes. These were equated to expert classifications where possible, for example *Restriction* was grouped with the various degrees of *Restriction* defined by the experts, and the *Obstruction* classifications grouped with that category. Two of the classifications, *Restriction with impaired gas exchange*, and *Obstruction – Irreversible* appeared to be compound classifications when compared to the expert knowledge base, consisting of both *Restriction* and *Impaired gas exchange*, and *Obstruction* and *No positive response to BD*. As such no direct comparison could be made.

Comparing Results

The comparison between student 1 and the combined expert knowledge base showed significant deviations, as expected. The student averaged 1.7 classifications per case with the majority (1831 cases, 62%) having 2 classifications. Some 266 cases were given no classification. However, holistic comparisons are likely to unfairly represent the student's knowledge. As the student only had the opportunity to examine 20 cases, there is a limited range of knowledge which can be acquired; whereas the expert knowledge base is compiled from over a hundred cases examined, and would be expected to be more complete for this reason alone. A measurement used to overcome this issue was to compare only classifications that appeared in both knowledge bases. Using this statistic, 192 (6.5% of the dataset) were found to be matches. That this is quite a small number indicates a large degree of incorrectness in the knowledge acquired from the student. Table 5-6 shows some of the generated comparison statistics for each corresponding classification grouping.

Classification Group	**Occurrences**		**Matches**	**Unique Occurrences**	
	Student 1	***Experts***		***Student 1***	***Experts***
Normal lung function	1746	1084	1060 (59.9%)	686	24
Restriction	2158	298	230 (10.3%)	1928	68
Obstruction – Mild	207	229	180 (70.3%)	27	49
Obstruction – Severe, irreversible	316	11	5 (1.6%)	311	6

Table 5-6: The results of the comparison between student 1 and the combined expert knowledge bases (percentages indicate the ratio of cases that match out of all cases identified in that class, by either expert or student)

The *Normal lung function* classification comparison shows a large discrepancy between the two, with the student overestimating which cases are normal. Although this does accurately identify 1060 of the 1084 expert-identified cases, including false positives the accuracy rate is 59.9%. An examination of the rules identifies

that the student uses none of the same conditions, or even the same attributes, as the experts. The most obvious difference is that the student does not include any volumes or diffusion components, focusing solely on spirometry. The definitions for *Restriction* are similarly disparate, with the ratio of matched cases roughly doubled, but still a very large number of false positives (1928) and a large number of false negatives (68). An examination of the rules indicates that the student again only used spirometry values to identify *Restriction*, whereas the experts exclusively used volumes measurements.

The comparison between *Obstruction – mild* definitions is much closer, with the student identifying 180 of the 229 cases correctly. However, their definition did also identify 27 false positives. Examining the rule definitions, the experts used $FEV_1/FVC < 0.7$ AND FEV_1 *% of predicted pre-BD* ≥ 80; whereas the student used FEV_1/FVC *% of predicted pre-BD* < 85 AND FEV_1 *% of predicted pre-BD* > 80 AND *FVC % of predicted pre-BD* > 90. This in itself identifies a likely discrepancy, based on the difference between the FEV_1/FVC percentage of predicted value and the explicit FEV_1/FVC ratio, and the one-sided inclusion of FVC percentage of predicted. However, the significance of those changes would not be obvious without the measured differences over the dataset.

The comparison between *Obstruction – Severe, irreversible* and the expert-defined *Fixed severe obstruction* is slightly misleading because of the severity component: as was seen in the expert knowledge base consolidation, the gradations appear not to be strictly defined but are rather subjective measures. However, the discrepancy displayed is far greater than could be reasonably attributed to that alone, with 311 false positives and 6 false negatives (55% of the cases correctly identified by the student's definition), indicating that although the rule conditions may appear reasonable there is a lack of understanding in some significant area.

Viewed Case Comparison

The same comparison when only considering the 20 cases seen by the student presents a similar result. The student averaged approximately the same number of classifications per case (1.8), and matched 2 cases to the experts' classifications (when considering classes used by both). The individual classification comparisons are summarised in Table 5-7; they show very similar ratios to those found by

comparing against the complete dataset. The only significant differences are the increased accuracy rates for *Restriction* and *Obstruction – Severe, irreversible*.

Classification Group	Occurrences		Matches	Unique Occurrences	
	Student 1	***Experts***		***Student 1***	***Experts***
Normal lung function	6	3	3 (50%)	3	0
Restriction	12	6	5 (38.5%)	7	1
Obstruction – Mild	2	3	2 (66.7%)	0	1
Obstruction – Severe, irreversible	7	2	2 (28.5%)	5	0

Table 5-7: The results of the comparison between student 1 and the combined expert knowledge bases, for the 20 cases seen by the student (percentages indicate the ratio of cases that match out of all cases identified in that class, by either expert or student)

Participant Feedback

The student described the system as difficult to use (2 out of 5 on the Likert scale) and that they would not use such a system again, citing the slow speed and glitches. However, the student described the cornerstone-based conflicting rule indications as helpful (4 out of 5). The student did feel that they learned more about lung function through the process, and indicated that while not willing to move on the scale from *2 (not confident)* they did feel more confident than before. The full details of all participant questionnaire feedback is summarised in Table 5-14.

5.4.2.2 Student 2

The second student participant was in the fourth year of a medical degree, and had recently completed a respiratory rotation as part of their training, working with a specialist respiratory unit. All significant bugs had been corrected and the system optimised, resulting in a much smoother interaction. It was noted that the student used more sophisticated terminology than others, as might be expected given a more advanced education and practical experience. It was also noted that the student showed less interest in the task the further it progressed, being content to accept classifications suggested by the system without a thorough examination of

the case. The student described themselves as neither not confident nor confident, with little experience with lung function reports, having seen 11-30 reports. The student also commented that the reports should use reference ranges rather than percentage of predicted values, and that beginners would be helped by the inclusion of a volumes graph.

Terminology and Equating Classifications

Perhaps due to their recent clinical experience, their use of terminology was much closer to that used in the expert knowledge bases, with more degrees of detail. The classifications defined by the student were: *Hypoinflation; Restriction*; *Mild Restriction; Moderate Restriction; Normal ventilatory function; Obstruction – moderate; Reversible moderate obstruction; Obstruction – severe, bronchodilators have a mild effect*; *Reversible mild upper airway obstruction; Mild diffusion impairment; Moderate diffusion impairment;* and *Diffusion impairment*. This raised some interesting points. The drawing of a distinction between *Hypoinflation* and *Restriction* is unusual.

As can be seen from the defined classifications, the student defined degrees of severity for each of *Restriction*, *Obstruction*, and *Diffusion impairment*, using the gradations *Mild*, *Moderate*, and *Severe*. However, no set of gradations were completed: for example *Restriction* included *Mild* and *Moderate* severities but not *Severe*, whereas *Obstruction* included *Moderate* and *Severe*, but not *Mild*. This is an apparent drawback from using a limited set of cases: the student did not see examples of cases which matched each of the criteria. The situation is further complicated by the subjective nature of each of these distinctions.

The incomplete range of severities makes equating the classifications difficult: based on the subjective nature of the number and thresholds of severities, grouping all severities together seems a reasonable choice. This is impractical however when some severities have been defined but others not, as there would be a distinct gap of coverage in the knowledge base missing those severities. Fortunately in this case the numbers and ranges of severities used seemed to roughly equate between the student and the experts, so the decision was made to equate individual severities where possible. Hence *Obstruction – moderate* was considered as a distinct classification rather than a member of the *Obstruction* group, as the student and

expert had both used this same term. This student also used some compound classifications presenting the same difficulties as student 1. The lack of singular counterparts for *Reversible moderate obstruction*, *Obstruction – severe, bronchodilators have a mild effect*, and *Reversible mild upper airway obstruction* make comparisons for these classifications more difficult.

Comparing Results

The comparison between this student and the combined expert knowledge base showed a similar pattern to the first student's comparison. Some 208 cases (7% of the dataset) matched when considering classifications used in both knowledge bases. This number itself is misrepresentative of the closeness of some of the definitions however. Examining the number of classifications per case shows that some of the discrepancy is likely a result of an incomplete acquisition of knowledge: 1432 cases (nearly 50% of the dataset) received no classifications in the student's knowledge base, with the majority of the other cases (1075) receiving one classification, giving an average of less than one classification per case (0.7). It is expected that this is partially a symptom of the specificity of some of the classifications, and partially that this student's interest waned as the process continued. Both indicate an incomplete knowledge acquisition process, suggesting that perhaps a wider range of cases need to be used, and highlighting the importance of having a student complete all cases to the best of their ability.

Classification Group	Occurrences		Matches	Unique Occurrences	
	Student 2	***Experts***		***Student 2***	***Experts***
Hyperinflation	567	568	567 (99.8%)	0	1
Restriction/Hypoinflation	374	298	298 (79.7%)	76	0
Mild diffusion impairment	310	496	129 (19.1%)	181	367
Moderate diffusion impairment	125	126	124 (97.6%)	1	2
Normal ventilatory function	177	1084	94 (8.1%)	83	990
Obstruction - moderate	115	545	61 (10.2%)	54	484
Moderate restriction	138	94	29 (14.3%)	109	65
Mild restriction	100	154	25 (10.9%)	75	129

Table 5-8: The results of the comparison between student 2 and the combined expert knowledge bases (percentages indicate the ratio of cases that match out of all cases identified in that class, by either expert or student)

Table 5-8 shows the comparison between classifications appearing in both knowledge bases. As alluded to earlier, some of these classifications show that although the overall matched cases measure indicates a very low level of similarity there is actually a strong agreement for some areas. The *Hyperinflation* classification, for example, is only different in one of 568 cases: examining the rules shows that the only difference is the use of *less than or equal to* by the expert and *less than* by the student, which was shown by the expert comparisons to be an arbitrary distinction. Similarly, *Moderate diffusion impairment* is only differentiated by 3 cases. The rules show that the student used the uncorrected D_LCO value for their definition, while the expert knowledge base used two rules, one using the uncorrected D_LCO and the other using the corrected D_LCO value, resulting in the difference. This would appear to be considered an unimportant distinction, as the two versions of the rule in the expert knowledge base were defined by the same

expert. The *Restriction/Hypoinflation* comparison is also quite similar, although the statistics indicate that the student used a broader definition than the experts. This discrepancy is due to the distinction between *Restriction* and *Hypoinflation* in the student's knowledge base. Interestingly, the student's *Hypoinflation* definition is identical to the experts' definition for *Restriction*; however their *Restriction* definition itself is markedly different and more complex.

Mild diffusion impairment however shows an unexpected difference, given the accuracy of *Moderate diffusion impairment*. Not only is there a large difference between the number of cases classified, the student's definition also classifies a largely different set of cases. This difference appears only to be due to the use of the uncorrected D_LCO value by the student and the corrected value by the experts. Both use otherwise identical conditions (*D_LCO % of predicted < 80* AND *D_LCO % of predicted > 60*), although the expert uses *less than or equal to* rather than *less than*.

The differences in *Mild restriction* and *Moderate restriction* are products of the very different definition for *Restriction* used by the student. The large discrepancy in *Normal ventilatory function* is due to the different attributes used to define the rule, and the evidently stricter ranges the student applies.

A further point of interest is the student's definition *Diffusion impairment*, as the cases classified by it overlap almost perfectly with the expert definition of *Severely impaired gas transfer* (*diffusion impairment* is a synonym for *impaired gas transfer*). The reason for this is the manner in which the student's rules were formed: the initial definition was for *Diffusion impairment*, specifying the rule *D_LCO uncorrected % of predicted < 80*. This was however superseded by the later definition of the exception rules for *Mild* and *Moderate diffusion impairment*, which specified the ranges *60 < D_LCO % of predicted < 80*, and *40 < D_LCO % of predicted < 60* respectively. This only leaves those cases with a D_LCO percentage of predicted below 40 to be covered by the initial rule. The significance of this is that it demonstrates the importance of clearly determining the detail of the classifications before the knowledge acquisition process is begun, particularly when using a limited set of cases as any mistakes such as this are less likely to be corrected by encountering one of the pertinent cases. While this definition could be considered incorrect when compared on a purely computational basis, the underlying statement could not actually be said to be false: those cases certainly do

exhibit *Diffusion impairment*. The only error is the lack of specificity equivalent to the other levels of specificity defined. As such this should perhaps not be considered as completely incorrect knowledge. The definition of the broad classification first, then severities later, provides a general fallback or ~~"s~~afety net" which ensures the relevant knowledge is applied even in cases which might be missed by the definition of more complex rules.

Viewed Case Comparison

The comparison results over the 20 cases seen by the student shows a marked improvement from the overall results in many areas. The student averaged 2 classifications per case, and matched 4 cases to the experts' classifications (when considering classes used by both), with a 22.5% average of classifications matched per case. The individual classification comparisons are summarised in Table 5-9; they show complete accuracy for *Restriction*, significant improvement in *Mild diffusion impairment*, and very significant improvement in *Normal ventilatory function* accuracy rates.

Classification Group	Occurrences		Matches	Unique Occurrences	
	Student 2	*Experts*		*Student 2*	*Experts*
Hyperinflation	4	4	4 (100%)	0	0
Restriction/Hypoinflation	6	6	6 (100%)	0	0
Mild diffusion impairment	4	5	4 (80%)	0	1
Moderate diffusion impairment	4	4	4 (100%)	0	0
Normal ventilatory function	4	3	3 (75%)	1	0
Obstruction – moderate	2	8	0	2	8
Moderate restriction	3	0	0	3	0
Mild restriction	5	6	4 (57.1%)	1	2

Table 5-9: The results of the comparison between student 2 and the combined expert knowledge bases, for the 20 cases seen by the student (percentages indicate the ratio of cases that match out of all cases identified in that class, by either expert or student)

Participant Feedback

This student described the process as fairly easy (4 out of 5 on the Likert scale). The cornerstone conflicts were again said to be helpful (4 out of 5), however the information additional statistics provided were described as unhelpful (2 out of 5). They described their confidence as unchanged and felt that they had not learned anything from the process. The student said they would use the system again however, if there were an initial set of rules already established in the system. All participant questionnaire feedback is summarised in Table 5-14.

5.4.2.3 Student 3

The third student participant was again in the third year of the medical degree. As such the student's knowledge was relatively shallow and incomplete, and the student indicated that they were only familiar with spirometry. The student stated

little experience with lung function reports (2 on the Likert scale), having seen less than 10 lung function reports before. They also described themselves as not confident (2 on the Likert scale). This was apparent in the definition of some rules, where the student could describe classifications for a case but was uncertain about how best to justify those classifications in a rule. A key example is an expressed desire to define a rule based on the shape of the flow volume loop, but being unable to as the student did not know which attributes represented those features of the graph. The uncertainty of definition also led to a high number of cornerstone cases conflicts being identified as the student defined new rules, and for the last five cases the student declared they were satisfied to define any vaguely plausible condition which would stop cornerstone conflicts.

Terminology and Equating Classifications

The classifications defined by this student were: *Normal lung function*; *Obstruction – mild*; *Obstruction – moderate*; *Fixed moderate obstruction*; *Obstruction – severe, bronchodilators have a mild effect*; *Mixed defect*; *Moderate mixed defect*; and *Severe mixed defect*. The use of the *mixed defect* terminology is rare in this study, having been identified early in development as a compound classification, and split into its component *Obstruction* and *Restriction* elements, with separate rules for each. Given that there are significant implications for having both conditions simultaneously, the experts did later include a rule explicitly reaching the classification *Mixed defect*. While having some understanding of what constituted a *Mixed defect*, the student expressed confusion about exactly what the components were or how they might be separately defined, suggesting that the student may have learned the pattern to identify a mixed defect but lacked the understanding of what it represented. As the experts made no distinction of severity of *Mixed defect*, and the student did not express which component the defined severities applied to, they were all grouped and compared to the expert-defined *Mixed defect*. The *Obstruction* classifications were compared against their equivalent severities where possible. The only classification with no direct equivalent was *Obstruction – severe, bronchodilators have a mild effect*.

Comparing Results

Considering only classifications appearing in both knowledge bases, there is a remarkable number of matching cases, with 1189 (40.2% of the dataset) meeting this criteria. The reasons for this are clearer when investigating the other statistics: the average number of classifications is below one per case, with the majority (2170, 73%) having one classification and 753 (25%) receiving no classifications. As shown in Table 5-10, 1046 of these single classification matches are caused by the closeness of definition for the *Normal lung function* classification. The high number of matching cases would therefore be indicative of the level of correctness of the student's knowledge, whereas the very small number of classifications per case shows that there are many areas in which the student has little or no knowledge.

Classification Group	**Occurrences**		**Matches**	**Unique Occurrences**	
	Student 3	***Experts***		***Student 3***	***Experts***
Normal lung function	1551	1084	1046 (65.8%)	505	38
Obstruction – mild	451	229	69 (11.9%)	352	160
Obstruction – moderate	91	545	68 (12%)	23	477
Mixed defect	113	61	14 (8.8%)	99	47
Fixed moderate obstruction	47	52	0	47	52

Table 5-10: The results of the comparison between student 3 and the combined expert knowledge bases (percentages indicate the ratio of cases that match out of all cases identified in that class, by either expert or student)

The other classifications and classification groupings show quite low percentages of matching cases and disparate numbers of cases classified. Investigating the rules show the differences in conditions; the student's rules are generally based on relevant or related attributes, showing an understanding of the underlying physiological effects, but a lack of knowledge of the clinical parameters in professional use. The large degree of difference between the end result suggest that

while the student's understanding may be grounded on some logical foundation, that understanding is not detailed enough to reach accurate conclusions.

Viewed Case Comparison

The comparison results over the 20 cases showed a surprising decrease in matches, with only 3 cases matching (15%). The average classifications per case increased to 1.1 however (2 cases had 2 classifications, the rest 1). The individual classification results were reasonably similar to those for all cases, showing very similar patterns, as shown in Table 5-11.

Classification Group	Occurrences		Matches	Unique Occurrences	
	Student 3	*Experts*		*Student 3*	*Experts*
Normal lung function	6	3	3 (50%)	3	0
Obstruction – mild	7	3	1 (11.1%)	6	2
Obstruction – moderate	2	8	2 (25%)	0	6
Mixed defect	3	4	0	3	4
Fixed moderate obstruction	3	2	0	3	2

Table 5-11: The results of the comparison between student 3 and the combined expert knowledge bases, for the 20 cases seen by the student (percentages indicate the ratio of cases that match out of all cases identified in that class, by either expert or student)

Participant Feedback

The participant described the software as fairly easy to use (4 out of 5 on the Likert scale). They identified the cornerstone conflict indications as very helpful (5 out of 5) and the statistical information as helpful (4 out of 5), saying that they used the statistical information for most rules (4 out of 5). The student made the comment that the program was effective and easy to use once they had become familiarised with it. They described that they were now more confident in interpreting lung function reports (4 out of 5), and that they improved their knowledge from the process, citing specific knowledge learned by seeing the statistics for how their

rules classified cases, and how their definitions conflicted. They indicated they would use the system again, stating that such a program would be useful in practical application on a ward, especially if the consultants used it and their knowledge could be used. All participant questionnaire feedback is summarised in Table 5-14.

5.4.2.4 Student 4

Student 4 was a fourth year student of the medical degree. They stated some experience with lung function reports (3 out of 5), having seen between 11 and 30 reports. They noted that they were unsure of their confidence (3 out of 5), stating that they ~~"~~know the basic principles but probably need to apply the knowledge in more situations". The student expressed throughout the process that they were concerned their classifications were not sophisticated enough, and remarked multiple times that although they were aware that they were not being assessed, it ~~"~~felt like a test".

Terminology and Equating Classifications

The classifications defined by the student were: *Evidence of gas trapping*; *Mild restriction*; *Mild airway obstruction*; *Moderately severe airway obstruction*; *Mild obstruction of small airways*; *Mild emphysema*; and *Moderate emphysema*. The *emphysema* classifications are difficult, as the experts did not go so far as to include diagnoses in their knowledge base, hence no analogue exists. As no other severities were defined for *Mild restriction*, and as it was found to be quite broad in scope with no minimum threshold, it was equated to the *Restriction* group of expert classifications. *Mild obstruction of small airways* was considered for grouping with the expert defined *Small airway obstruction*, but ultimately was excluded as the student's definition was very specific to the *mild* component and a full comparison would be uninformative. The *Obstruction* classifications were compared directly with their counterparts.

Comparing Results

Using only shared classifications, 135 cases (4.6%) find the same results. This student displays a lower breadth of knowledge than previous participants however, with an average of 0.4 classifications per case, 1815 with no classifications and 1148 with a single classification. Examining the rule structure it appears that this is

not due to overly specific rules, but to the definition of overly general rules: the initial rules defined were very broad, utilising conditions such as *FEV_1/FVC % of predicted < 100*. When these were found to classify further cases incorrectly, the student added an exception to the rule to define a new and separate classification, when the appropriate procedure would have been to remove the existing classification (adding a stopping rule), and then add a new classification (adding a new, non-exception rule). While not exclusively defining rules in this format, the nested nature of many of the rules ensured that no case would receive more than one classification.

Table 5-12 shows the classification comparisons. The definition for *Restriction* is representative of the student's definitions in general, with the conditions *TLC % of predicted < 100 AND FEV_1/FVC % of predicted < 100*. These conditions are clearly far too general, as the use of *< 100%* would incorporate even cases that have a measurement 99.9% of the predicted value. These definitions indicate the level of the student's knowledge: they were following the heuristic ~~"a~~ case is restricted when TLC is reduced", which agrees with the experts' definition, but the student had no understanding of what constituted a significant reduction. This would seem to be an example of a student learning a pattern without understanding the pattern. The definitions for *Evidence of gas trapping*, *Moderately severe airway obstruction, Mild emphysema* and *Moderate emphysema* exhibit the same problem, although *Mild airway obstruction* uses the more realistic conditions *FEV_1/FVC % of predicted < 80 AND FEV_1/FVC % of predicted > 60* and finds an approximately equivalent number of classifications to the expert, although this is also shown to be a very different set of cases. In addition to being overly general, *Mild emphysema* exhibits the opposite problem with the condition *FEV_1/FVC % of predicted = 100*, which actually serves to balance out the frequency of classification, although it classifies incorrectly.

Classification Group	Occurrences		Matches	Unique Occurrences	
	Student 4	*Experts*		*Student 4*	*Experts*
Restriction	474	298	129 (20.1%)	345	169
Mild airway obstruction	131	137	26 (10.7%)	105	111
Moderately severe airflow obstruction	304	102	20 (5.2%)	284	82
Evidence of gas trapping	98	405	15 (3.1%)	83	390

Table 5-12: The results of the comparison between student 4 and the combined expert knowledge bases (percentages indicate the ratio of cases that match out of all cases identified in that class, by either expert or student)

Viewed Case Comparison

The comparison results over the 20 seen cases showed 7 matches (35%) using shared classifications, with exactly one classification per case. *Restriction* showed a large increase in accuracy, but still with 2 false positives (out of 7 positives) and 1 false negative (out of 6 true positives). Full numbers are summarised in Table 5-13

Classification Group	Occurrences		Matches	Unique Occurrences	
	Student 4	*Experts*		*Student 4*	*Experts*
Restriction	7	6	5 (62.5%)	2	1
Mild airway obstruction	2	3	1 (25%)	1	2
Moderately severe airflow obstruction	4	2	0	4	2
Evidence of gas trapping	2	6	2 (33.3%)	0	4

Table 5-13: The results of the comparison between student 4 and the combined expert knowledge bases, for the 20 cases seen by the student (percentages indicate the ratio of cases that match out of all cases identified in that class, by either expert or student)

Participant Feedback

Student 4 described the system as very easy to use (5 out of 5). They were however unsure of the usefulness of the cornerstone conflict indications (3 out of 5). While stating an understanding of the rule statistics presented, they were unsure of their usefulness (3 out of 5) and they never influenced their rule making decisions (1 out of 5). The student indicated no increase in confidence, but that they had learned from the process, stating that the system had identified multiple areas to ~~r~~ead up on". The student said they would use the system again. All participant questionnaire feedback is summarised in Table 5-14.

	Student Participant				
	1	**2**	**3**	**4**	**Mean**
Experience with lung function reports	3	2	2	3	2.5
Number of reports seen	1-10	11-30	1-10	11-30	-
Confidence in interpreting lung function reports	2	3	2	3	2.5
Ease of use of software	2	4	4	5	3.75
Usefulness of cornerstone conflict indications	4	4	5	3	4
Usefulness of statistics in rule definition	3	2	4	3	3
Influence of statistics on rule definition	NA	NA	4	1	2.5
Confidence in interpretation post-test	2	3	4	3	3
Increase in confidence after using system	0	0	2	0	0.5
Did participant learn from the process?	yes	no	yes	yes	-
Would participant use the system again?	no	yes	yes	yes	-

Table 5-14: Summary of pre- and post-acquisition questionnaire answers (NA indicates not answered)

5.5 Discussion

5.5.1 Knowledge Consolidation

The successful identification and resolution of conflicts, and the successful compilation of multiple experts' knowledge into the expert system presented in Chapter 3, indicate that this approach can be used for knowledge consolidation. The results presented here provide details of how well the method functions, what deficiencies exist and how it might be improved.

5.5.1.1 Equating Classifications

The identification of classification equivalencies is one of the weak points of the method presented here, as it required some manual analysis work and post knowledge acquisition communication between the experts involved. The extent of this problem is dependent on the domain and the experts involved; in domains in which the terminology is clear, consistent, and unambiguous, no classification equating would be necessary. The domain used in this study showed significant differences in the terminology used for the same classifications, not only between different experts but at times between the same expert at different times. It is worth noting that in this study no participant ever expressed confusion about the meaning of any particular terminology. Even student participants who demonstrated a lack of understanding of some classifications seemed familiar with the terminology used, and indeed in this study the definition of classification equivalencies was not considered to be a difficult task to complete manually. Nevertheless, it is necessary in order to computationally compare results accurately; the doubling in average matched classifications after equating classifications, and the drop of unmatched classifications from 52 to 28, show how much extra knowledge was able to be compared between the expert knowledge bases. The minimisation of the effort required for this task is a problem to be solved, especially if considering applying the method to domains whose terminology may not be as easy to resolve.

There are two specific problems related to terminology identified during this study that require resolution: the use of compound classifications and the definition of differing levels of severity both present difficulties for comparison. The major problem with the definition of severities is that different experts used different

numbers of grades, and different thresholds to define each grade, making direct comparisons at times impossible. The definitions are clearly subjective, and do not necessarily conflict with each other: in most situations each expert would be satisfied to accept a slightly more or less detailed scale of distinction. However ignoring the defined scales loses some of the detail in the acquired knowledge, and risks missing some genuine conflicts of opinion. For these reasons the solution used in this study was to compare gradations directly where possible, but otherwise to group them together and compare holistically, as this would at least allow some comparison of the knowledge defined. This is a workable solution, as shown by the effective comparisons made here, but other options may be considered.

The use of compound classifications presents other difficulties. This problem was identified relatively early in expert knowledge acquisition, as experts sought to define the classification *Mixed defect* (*Restriction* and *Obstruction*), and it was advised that experts separate classifications wherever possible. This was largely adhered to, although a common exception was the definition of various combinations of *Reversibility* or *Positive response to bronchodilators* with *Obstruction*. This was not a significant problem in comparing and consolidating the expert knowledge bases, as each expert generally seemed quite able to follow the recommendation of defining singular classifications. The occurrences of compound classifications such as *Reversible obstruction* were relatively easy to compare in this study by examining the rules and identifying the conditions which corresponded to each component. The alternative approach to breaking down the compound classification is to identify the component classifications as defined in the other knowledge base, and identifying those cases which match both classifications. This seems to the author to be a practical and simple solution for more complex situations.

There are many approaches to avoiding or improving the definition of classification equivalencies. The use of a method such as formal concept analysis to identify conceptual differences, as described by Richards and Compton (Richards, 1998; Richards & Compton, 1997c), would likely improve this process by providing a visualisation of the classifications and rule conditions in each knowledge base. Depending on how different the definitions are, this could assist in identifying equivalent terminologies based on similarities of conceptual structure. The

hierarchical separation of classifications and rules into conceptual components may also assist in the general comparison of compound classifications. A discussion on the comparative effectiveness of formal concept analysis and the method presented here is provided shortly.

There are other options for improving the classification equating task. One of the simplest alternatives is to perform a level of domain modelling before the main knowledge acquisition, and then be prescriptive about the classifications that can be used. This would not in fact remove the problem of identifying which classifications are equivalent, but would rather shift the work to an earlier stage; although it might be expected to be an easier task as idiosyncrasies and terminological differences are identified and resolved before they become widely manifest in the knowledge base. This would also resolve any issues with differing gradations of severities and the definition of compound classifications. This approach was rejected from this study however, for a few significant reasons. Firstly, it restricts the range of knowledge that can be acquired and expects experts to list all knowledge that they might define before they have seen any cases, making the uncovering of tacit knowledge a more difficult task. Secondly, there can be a loss of detail in the knowledge acquired as the definitions will often be more generic than otherwise in order to accommodate a wide range of experts' views; and as experts seek to conform to those generic classifications they will not be producing their usual output. Thirdly, the differences in terminology between experts can be an interesting result in itself. Lastly, and most practically, this study did not have a limit on the number of experts that would be invited to participate, and it seemed undesirable to define a prescriptive set of classifications for experts to use without their input into what those classifications should be.

There are other options for improving the terminology equating process. While not particularly necessary for this domain, the automatic identification of similar classes could be useful in domains where differences in terminology present a significant problem. The methods presented in this study can already find similar classes based on the attributes of the class membership, and it is expected that this could provide a simple way of suggesting equivalent terminology for experts' knowledge bases. This method would not replace the classification analysis, but should alleviate some of the work involved.

5.5.1.2 Quantified Comparisons and Conflict Identification

One of the most significant benefits of this approach is that it does not rely solely on abstractions and theoretical understanding to compare knowledge. The application of each knowledge base over a dataset allows a measure for how significant each difference is, and to quantify the overall comparison in a number of ways. The frequency of the number of classifications made by each knowledge base (as shown in Table 5-4: Frequency of number of classifications per case, for each knowledge base), and the overall numbers and ratios of matches, can provide a general impression of the relative qualities of each knowledge base. This impression is then qualified by the numbers of cases given each individual classification. For example, the higher frequency of larger numbers of classifications generated in this comparison indicated that the individual expert knowledge base used more general classifications, and the individual classification numbers identified which classifications those tended to be.

The quantification of the individual differences is important in assessing the importance of conflicts. A subtle difference in definition that may appear to be inconsequential might in practice have a significant difference on the end result; for example, the 450 case (34.8% of all cases identified) difference between definitions for *Diffusion impairment* are entirely due to the interchanged use of the corrected and uncorrected values for D_LCO, with the rules being otherwise identical. Conversely, a difference that may appear to be significant may have quite a minor practical effect, as for example with the definitions for *Normal lung volumes*: one knowledge base made use of TLC and FRC in defining rules, whereas the other used at times TLC and other times only RV, and yet the differences were quite minor with an 84.8% agreement overall.

The quantification of the differences between definitions has further importance to knowledge base consolidation because of the nature of the data. While there do exist standards as to how to clinically determine different classifications, the inter-related nature of each of the attributes mean that there are typically multiple methods to define a classification, using different attributes, all based on reasonable underlying principles. For example, as described in section 4.2.3 there are many definitions for the classification *Reversibility* that are in clinical use and much disagreement over the best definition to use (American Thoracic Society, 1991;

Borg, et al., 2004; Jenkins & Young, 2004). This is supported by the knowledge acquired from the different experts in this study, as the two specialist experts entered differing definitions for a number of classifications that required resolution. Two of these conflicts were only resolved by examining the number of cases affected by each different definition, after which the experts were able to select the definition that best fit their expectations and understanding. These two examples best demonstrate the benefit of being able to quantify the difference between two definitions, where a conceptual comparison was simply insufficient. Those two examples also highlight that even experts cannot always be sure what the impact of differences in definition will be, as they requested the number of cases each classified by each definition. The identification of which differences are significant and which are inconsequential can also reduce the number of conflicts which need to be brought to the experts' attention, an important consideration given the value of expert time.

Formal Concept Analysis

As described earlier, formal concept analysis also provides a means of comparison between the knowledge defined in each knowledge base, and as such overlaps with the goals of this method. Where the approach presented here differs is in the identification of quantified results and exemplars of differences, as will be discussed. Whereas formal concept analysis can better provide a visualisation allowing the identification of significant conceptual differences, this approach is more able to identify subtler differences, and importantly, the significance of those differences to the end classification. It is suspected that this approach would also perform better for larger knowledge bases, where a visualisation would be too large to be easily inspected. However, this method is not intended as a replacement, but rather presents certain comparison information that other approaches do not: it is suggested that a combination of the two methods might provide better results, depending on the domain and knowledge structures. Further study is needed to qualify this conjecture however.

5.5.1.3 Conflict Resolution

As has been described, quantifying the differences between different definitions of a classification can assist in resolving those conflicts. While most of the conflicts

encountered were resolved by simply viewing the different definitions and identifying a sensible solution, or presenting the conflict to the experts who could immediately resolve it, some of the conflicts required the use of the classification statistics. These methods resolved all conflicts that were found between the experts in this study; however, the comparison method provides a further option for conflict resolution should it be necessary.

For a more rigorous conflict resolution process, it is suggested that as part of the resolution discussion each expert could be presented with a case or cases that are exemplars of their conflict. These cases are easily identified by the method used: the interface as developed provides the option to view the set of cases described by any particular statistic, for example those cases which uniquely have the classification *Obstruction* in the first knowledge base. The presentation of the cases, with the classifications added by each expert, provides the full context in which the conflict should be considered. Examining the set of cases that are causing a particular conflict can also allow the identification of other attributes which may differentiate the groups, or further refine their definitions. This should ensure as little confusion as possible and may elicit more detailed knowledge, particularly tacit knowledge, which may not have been forthcoming in a less specific context. This does however require the availability of the experts to respond to the conflicts in a detailed manner, which is not always possible, as is demonstrated by this study.

Another area the method might be expanded is in the automatic identification of why two definitions produce different results. The method as described generates a measurement of how different two definitions are, and it provides the rules as explanation; however, it does not automatically identify what the differences in definition are, nor the significance of each of those differences. It is expected that an automated analysis of the effects of each differing rule condition on the results could identify this relatively easily, but it is unknown whether the benefits would outweigh the cost of computation.

5.5.2 Teaching and Learning

In addition to being used to consolidate two expert knowledge bases, the knowledge acquisition and comparison method was also tested as a general tool for education and assessment. It was hypothesised that the knowledge acquisition and subsequent

comparison would provide practical experience and help participants explore their understanding of lung function, and that the knowledge comparison may also provide a tool for the assessment of participants' knowledge against the consolidated expert knowledge base.

5.5.2.1 Practical Experience

The process of acquiring a participant's knowledge in the structured ripple down rules format, with associated statistics for how that knowledge applies over a dataset, showed some promise as an educational tool. All but one student indicated in their questionnaires that they felt they had learned from the process, even without any feedback to compare their definitions to the experts'. One student indicated a significant increase in confidence (from "not confident" to "confident") in interpreting lung function reports, describing specific knowledge learned from the process. Another student described an increase in confidence, but not significant enough to move them from "not confident". Likewise all but one student participant indicated they would use the system again, the one exception being the first student who cited the initial bugs as the only reason for their reticence. As there was no instruction of any sort about lung function knowledge, no expert knowledge was provided, and the administrator of the system made it clear that he did not have any education in lung function, the described gain in confidence and knowledge learned is assumed to have come from the practical experience of working with lung function tests and seeing the practical application of learned theory. This is supported by the participants' comments: two students commented before beginning that they had some understanding of principles, but felt they needed to apply the knowledge in practical situations. The two students who described an increase in confidence both attributed this to the practical application of theory; another student identified multiple areas to "read up on", based on shortcomings identified because in attempting to apply knowledge they discovered gaps and inconsistencies. This is an expected result as the benefits of practical application of theory are well-described, with the situated learning and constructivism schools of thought advocating the necessity of practical application of theory in learning (Anderson, 2004; Duffy & Cunningham, 1996; Lave & Wenger, 1991). Nevertheless it is an important result that the students learned through the process and were supportive of the approach.

The impact of the dataset statistics provided in the rule definition process appears to be minimal, with participants averaging 3 out of 5 on the Likert scale for usefulness, and a 2.5 for influence on rule definition. It might be expected that the students would not be able to take full advantage of the statistics in guiding their definitions, as to do so requires at least a reasonably confident expectation of the frequency of appearance of each classification, and a reasonably confident expectation of the ranges displayed for each attribute. Given the self-described low level of experience, knowledge and confidence for the students involved, it is perhaps unsurprising that they did not devote much attention to the statistics. As the only student with practical experience, the second student participant commented that the statistics were a good and useful idea once they had been explained to them. However, they did not use that knowledge at any point – whether that was because of a lack of understanding of how to use it, a lack of interest, or a lack of time, is unclear. A similar pattern was noted for the other participants, with the one exception being student 3, who did examine the statistics for each rule defined and described information learned from doing so. Student 3 noted in the post-acquisition questionnaire that the statistics indicated the percentage of cases present for the classification being defined, and that they used that information to adjust the rule conditions to avoid making classifications that were too broad. At the very least, the student learned from this what sort of distributions to expect from some lung function attributes, and how better to differentiate opposing classifications. This demonstrates some usefulness to making the statistics available, and the results demonstrate no negative impact; although there is the possibility that the figures may confuse and intimidate users, and that improved interface responsiveness by removing the statistics may be of benefit. A further negative is that only one student gained any benefit from the statistics; but this was also the only student who seemed to show any interest in them. There is no evidence to show a benefit from removing the statistics, but some positive results were achieved from them; the matter requires further study.

The benefit of the rule conflict indications, based on the identification of cornerstone case conflicts, is clearer. The average questionnaire result described the usefulness of the conflict indications as 4 out of 5 (–Useful"), with one participant being uncertain of the impact and one student describing them as very useful. It was

noted that conflicts were quite common, with every participant defining conflicting rules, which forced them to analyse both the accuracy of their knowledge and analyse the conflicting cases to identify relevant differences.

5.5.2.2 Knowledge Comparisons

The comparison of participant knowledge to expert knowledge can provide a variety of insights into the nature of that participant's knowledge. Of particular benefit is the identification of specific weaknesses, gaps in knowledge, or misunderstandings.

Identifying Weaknesses

Terminology and Classification Equivalencies

The terminology used in each knowledge base can provide insights into the knowledge expressed, particularly when compared with the experts' terminology. These insights are revealed in the process of finding the classification equivalencies in the knowledge bases. As expected, each participant defined much fewer classifications than were present in the expert knowledge base. The classifications that are defined by each participant give the simplest indication of the student's level of knowledge, or at least the knowledge that they feel confident expressing. For example the first student defined the classifications *Restriction*, some severities of *Obstruction*, irreversibility, and *Normal lung function*, indicating a lack of knowledge of any volumes- or diffusion-based classifications. This is valuable in itself, although it does not describe how correct their knowledge is. Equating the terminology used can also reveal confusion on the part of the student: for example the second student defined separate classifications *Restriction* and *Hypoinflation*, a distinction not made by the experts.

Many of the classifications defined by the student participants also highlighted the problems that the method faces with compound classifications that cannot be directly compared. This was a more common problem with student participants than with experts, for example, the first student's classifications *Restriction with impaired gas exchange*, and *Obstruction – Irreversible* appear to be compound classifications when compared to the expert knowledge base. This is perhaps due to the shorter acquisition process, with less instruction as to how and what type of

classifications to define, and perhaps as the participants had less knowledge to be able to separate and differentiate compound classifications. For example the third participant defined *Mixed defect* (which the experts separated into *Restriction* and *Obstruction*), but when it was suggested that they define classifications as individually as possible they professed that they did not know how to separate the classification. For these reasons the comparison of the compound classifications is more of an issue for student-expert comparisons. In some instances the conditions of the compound classifications may be separated into two rules to allow direct comparisons, but to ensure accuracy this would require consultation with the person that defined the rule. Whether a discussion was attempted or not, there is still the possibility that the classification is not in fact a compound classification, at least to the person defining it: that the definition uses characteristics of the case that are only present when both classification components are present at the same time, and therefore cannot be reduced to separate rules. In this situation it can only be compared against a combination of the individual expert-defined classifications, if a comparison is to be made.

Comparison Results

Although the terminology can describe general limits of knowledge and suggest flaws, these are relatively meaningless without some identification of the correctness of the knowledge that is there and the significance of the problems. The definition of *Restriction* and *Hypoinflation* separately is not in itself incorrect, as the experts also used both terms at times. It is only once the results of the classifications across the dataset are compared to the experts' results that the discrepancy is revealed. The large difference in number of cases identified in this situation highlighted the flawed knowledge. Similarly the comparisons provide measurements for how accurate each classification is, including areas in which the student has a strong understanding. The strong correlation between cases classified for the *Hyperinflation* and *Moderate diffusion impartment* classifications between the second student participant's knowledge base and the experts knowledge base clearly shows the students' strengths The quantified results for the first student's definition of *Obstruction*, revealing 180 of 229 cases correctly classified, with 27 false positives, gives a clear and unambiguous measurement of how correct the student's knowledge is.

As with the expert comparisons, the significance of a difference in definition is not always obvious. A subtle difference in definition might in practice have a strong or limited effect. For example, one student's use of the uncorrected D_LCO value rather than corrected showed a very significant difference for *Mild diffusion impairment*, despite the seeming insignificance of the difference. This highlights those flaws in knowledge that have an empirically more substantial impact on the end result, rather than relying on an intuitive sense of significance, which might be incorrect.

As with the expert to expert comparisons, a holistic comparison can also provide useful information, such as general breadth of knowledge, general accuracy of knowledge and student confidence. The breadth of knowledge can be indicated by the number of classifications per case: for example, the knowledge base developed by student 3 showed 73% of cases had one classification and 25% had none, demonstrating a quite narrow range of knowledge. The number of matching cases however showed 1189 or 40% of the dataset, when considering classifications present in both knowledge bases; indicating that although there may not be a broad range of knowledge, the knowledge which is there is reasonably accurate.

Evidence-based Resolution

As was suggested for expert comparisons, the use of exemplars is suggested as a strong basis for demonstrating weaknesses in knowledge. Once a problem in knowledge is identified, the expert's knowledge can be presented along with a case demonstrating the difference in action. It is suggested that the presentation of a real example for the student to consider would allow the student to not only examine the attributes present in the expert's definition, but also allow them to examine the associated pattern of other related lung function variables. This would be expected to further reinforce their understanding of the classification.

Assessment

Given that this method quantifiably compares the knowledge of a student to a more reliable source, a logical application of the method is for student assessment. A number of considerations apply to this however. Although the method provides some general measurements of accuracy, the results here show that none can be considered individually sufficient as an analysis of student knowledge. The overall

accuracy measure reached 40% for student 3 and 7% for student 2; however, when considering the accuracy of individual classifications, and the more logical rule definitions provided by student 2, it is clear that this student had a better understanding of lung function. The reason for student 3's higher score is evident in the individual classification statistics, with the single classification *Normal lung function* providing 1046 of the 1189 correct matches: as the student happened to define the most common classification accurately, their overall accuracy is higher. As pointed out previously, this accuracy measure needs to be supplemented by considering the breadth of knowledge, indicated in part by the number of classifications per case. However, this also fails to differentiate between the two students to any significant degree. A more useful measure in this situation is the average accuracy for each classification defined: this gives 19.7% for student 3, and 42.5% for student 2. While this does give a good estimate of the accuracy of each student's knowledge base, it should only be considered as one component of any assessment.

The method could still provide useful assistance to an assessor however. The quantification of how many cases a classification covers has been shown to be useful, with experts unable to predict what the results of differing definitions will be over a dataset. This then should also be a useful tool to an assessor in identifying the accuracy of a definition. Further advantages are provided in the identification of specific areas that a student has difficulty in, and identifying specific cases that highlight those difficulties, presenting not only a method of assessment but also the means to improve the student's knowledge.

Student Knowledge Acquisition

There are some concerns about the accuracy of the knowledge acquisition process, and the impact it may have on the knowledge acquired. These concerns relate to the restricted set of cases used, the complexity of the process, and the relative inexperience of the participants.

The knowledge acquisition process performed with the student participants was necessarily modified from a typical MCRDR approach. A significant change was that the number of cases was restricted to 20, rather than the usual approach of allowing knowledge acquisition to continue for as many cases as needed for the rate

of corrections to plateau. The biggest concern with this change is that there would not be enough cases to allow the participants to express their knowledge, and to identify and correct mistakes. The lack of complete knowledge acquisition is suggested by some of the results: the knowledge base developed by student 2, for example, contains definitions for *mild* and *moderate Diffusion impairment* and *Restriction*, but not *severe*, whereas *Obstruction* had definitions for *severe* and *moderate* but not *mild*. This particular missing knowledge is not a serious concern, as the missing definitions can be derived from the others if necessary; in fact the general definition *Diffusion impairment* effectively represented the student's definition for *Severe diffusion impairment* by a process of elimination. The missing terms are however indicative that the knowledge acquisition may be insufficient. This is supported by the results for the 20 seen cases compared to the overall results. Each participant defined some classifications which were reasonably accurate for the 20 cases examined, but which then became much less accurate over the complete dataset. Student 2 in particular defined some very accurate rules for the reviewed cases, with 3 classifications showing 100% agreement with the experts. Although the accuracy of these definitions dropped over the full dataset, it might be expected from this that if the student had seen some of those incorrect cases they could have added further exceptions and improved their definitions. The other participants displayed this to varying degrees, although no pattern could be discerned to predict which definitions would extrapolate well and which would not. The difference in mean classifications per case, between seen and unseen cases, is also significant. While this would seem to indicate that the students had more knowledge that was not acquired, part of the discrepancy comes from the students not having a strong understanding of the underlying patterns. The lack of understanding is shown by the inaccuracy of some of the rules, even over the reviewed cases, and it results in the definition of rules which are based too heavily on specific attributes of the current case, rather than being expressions of an underlying pattern.

Despite some evidence that the students had more knowledge to be acquired, there are however a number of factors ameliorating this concern. The 20 cases are still thought to have provided sufficient breadth of classifications. The cases were selected with expert consultation to find a spread of cases that provided multiple

examples of each of the classifications the students were expected to display. Some cases were chosen as typical examples of a classification, and others were selected as difficult borderline cases to hopefully derive detailed knowledge. The relative inexperience of the participants assisted in ensuring that a limited set of cases could cover the relevant areas. Furthermore, the presentation of statistics for how their definitions applied over the full dataset provided some access to the much larger store of data, to help elucidate detailed and accurate knowledge. Nevertheless, in any similar study or application, the number of cases used should be carefully considered to balance the time required with the data needed to express the participants' knowledge. The other impact of reduced cases in the MCRDR acquisition process is that it may not provide sufficient data to verify previously defined rules, as cases exemplifying previous errors may not be present. This is resolved somewhat by the addition of dataset statistics, but it is argued that this is not an especially relevant concern in this situation: the goal is to acquire the student's knowledge for comparison and assessment, rather than ensuring that all knowledge is complete and correct. If the participants add faulty definitions, they would presumably be relying on that same faulty knowledge in a practical situation, which is exactly the kind of mistake that the method is hoping to identify. Classifications that were shown to be completely accurate for the reviewed cases might be considered to be nominally correct, given there is no evidence that the student could not correct their classification when presented with a problematic case. However, any classification which incorrectly classifies, or fails to classify, reviewed cases should certainly be considered as evidence of a weakness in knowledge; and by applying that weakness over the larger dataset a more complete understanding of the significance of that flawed knowledge can be found.

The problem of an incomplete knowledge acquisition process was raised in the results, both as a consequence of an insufficient range of cases, and for one participant due to fatigue or lack of interest. It is suggested that the problems presented by this latter point would not apply were the approach developed as an assessment tool, as this would provide a more meaningful outcome to the participants. A similar improvement might be seen if developed specifically as a learning tool. Developing a process that participants can and will complete to the best of their ability is a significant concern for further research in the area, and

financial incentives may be necessary to encourage participants to engage in future research projects that will not directly affect their real grades.

The complexity of the knowledge acquisition process is also a concern. In particular, the results show some problems introducing the participants to the method in a short time frame, including how to define rules and the types of terms they should use. The students were given guidelines for how to define their classifications, such as to try and make general classes then make more specific ones, and not to use compound classifications but to separate classifications into individual components wherever possible. Apart from this they were largely left to perform the knowledge acquisition as they saw fit, with no prescription of specific classification terminology. This choice made for some difficulty in analysis but was expected to better represent the student's knowledge, and provide more realistic results of the application of such a method. If this method was applied to a class of students, for example, many of the same problems would be encountered as each one cannot be supervised individually. As would be expected, difficulties were encountered as students defined compound classifications or used differing paradigms of classification than expected. For example, one student went beyond simple clinical interpretation and attempted to define a diagnosis of *Emphysema*. Further difficulties occurred with students misunderstanding the rule structure, a problem also encountered with the experts in this study: student 4 encountered difficulties after defining overly broad rules, then only defining exceptions when the rule produced incorrect results, without defining any new rules or stopping rules. The third student also remarked that the system was effective and easy to use once they had become accustomed to it. These problems make it clear that were any such system to be employed, significant consideration needs to be made to present the options unambiguously and with clear instructions. However, given that the comparisons in this study were successfully made and weaknesses identified, and given the positive feedback from the participants both in terms of learning and willingness to use the system again, the method seems to show some promise for further research. Important areas of research include determining the effects of being prescriptive about the classifications used, and whether the process can be explained and presented in a sufficiently understandable manner.

The final concern is with the suitability of the participants. The evidently large discrepancy between the knowledge of the student participants and the experts' knowledge calls into question the usefulness of the comparisons. This is perhaps a failing of the experimental design; however, the results still serve to demonstrate the function and efficacy of the method. There is also no reason why future studies could not develop a knowledge base tailored to a specific knowledge level, such as the knowledge expected from a certain class or professional, to provide a more relevant comparison. It has been considered that as the students' knowledge is quite shallow and relatively undeveloped, they perhaps should be assessed in stricter terms of correctness of method rather than focusing on similarity of end result. However, the process has been shown to be an effective learning tool for these students, and provides real practice and application of theory which they were apparently otherwise lacking: the third year students had seen less than 10 cases each and the fourth year students only between 10 and 30, so working through another 20 cases should almost certainly provide useful reinforcement of learning. The approach as presented also allows an identification of the gaps in knowledge of each participant, which is of some importance. Some of these students did, after all, go on practical hospital rotations and at times may well be expected to put their knowledge to the test in a real situation. Regardless of the relative merits of this exact method to the particular participants, the results still serve to demonstrate the application of the method in effective comparisons of knowledge.

5.6 Conclusions

This chapter described a method for quantifiably comparing the knowledge of multiple experts. This knowledge is acquired by a ripple down rules knowledge acquisition process, and the resultant knowledge bases compared over a dataset in order to identify and reconcile conflicts. The results of comparing and consolidating two expert knowledge bases in lung function were presented: they showed the ability of the method to identify important conflicts between experts' knowledge, and to provide quantified evidence on the differences between definitions to assist in resolving those conflicts. This not only provides a method that can consolidate the acquired knowledge of multiple experts, improving the knowledge acquisition outcomes, but addresses the issue identified earlier in this study of finding

resolutions to knowledge conflicts that are acceptable to all parties and improve the knowledge of the experts involved.

The chapter also described the potential application of this comparison method as a teaching and learning tool, and presented the results of comparing knowledge bases acquired from four medical students with the combined expert knowledge base. Participants indicated that they benefitted from using the method by identifying weaknesses in their knowledge, by learning from the dataset statistics provided, and by gaining practical experience in examining and applying theory of interpretation to lung function reports. Participants also indicated they would use a similar system again. This provides some quite positive results for application as a learning tool, despite very little focus on participant education. The potential for future development of this method, and its application as an assessment tool are discussed. Although there are measurable benefits to using the approach, more work needs to be done to ascertain how this method might be used to produce an assessment rating.

The approach to comparison is not expected to replace conceptual knowledge comparison methods such as formal concept analysis, but rather complement these techniques. Some concerns exist with the problem of differing expert terminology, and with ensuring that participants understand and conform to the procedures of the knowledge acquisition. However, it is concluded that the method can effectively quantify differences in knowledge, identify significant differences, and assist in their resolution. The method can also assist in learning through the provision of practical experience, and provide a measure for how correct participant knowledge is in comparison to defined expert knowledge.

Chapter 6 Summary

This thesis examined the issue that knowledge discovery methods often lack the ability to incorporate existing domain knowledge (Sinha & Zhao, 2008), and that this omission makes knowledge discovery in complex domains impractical (Adejuwon & Mosavi, 2010; C. Zhang, et al., 2009). Although attempts have been made to resolve this problem, studies frequently identified that in order to be successful, these methods require a knowledge acquisition or knowledge engineering process that is still impractically expensive (Kotsifakos, et al., 2008; Liu, et al., 1997; C. Zhang, et al., 2009). Based on these requirements, a method was developed to overcome this problem based on the MCRDR knowledge acquisition approach. This method was tested in the suitably complex domain of lung function (Cios & Moore, 2002a; Roddick, et al., 2003). In developing the method it was noted that the base MCRDR approach was not able to take advantage of a dataset to assist in knowledge acquisition; hence an enhancement was developed to provide additional evidence-based validation. It was also noted that existing methods for collaborative knowledge base development (Richards, 2009; Vazey & Richards, 2006) or knowledge base integration (Beydoun, et al., 2005) lacked an ability to assist in conflict resolution, and lacked a focus on improving the experts' knowledge; therefore a comparison and consolidation method was also developed and tested.

Several findings have been presented in this thesis, from a range of experiments. The first experiment, presented in Chapter 3, described the development of a knowledge base for the field of lung function that is capable of interpreting patient lung function test results. This knowledge base was developed through a modified MCRDR method: the knowledge was acquired from multiple experts, both collaboratively and through post-acquisition consolidation; and a large dataset of cases were used to provide additional validation of acquired knowledge. The effects of these modifications were examined: the use of multiple experts seemed effective, but was inconclusive without an effective comparison and without a more extensive evaluation, particularly for the collaborative knowledge acquisition. The validation modifications, while likely successful in improving the end result, seemed to

complicate the acquisition process by shifting experts from a case-based to a rule-based focus with some detrimental effects.

The second study presented in this thesis, described in Chapter 4, was a method for the incorporation of complex domain knowledge into a knowledge discovery process. The method applies the consolidated knowledge base, together with common data mining techniques such as association rule mining and information gain comparisons, to the analysis of a dataset. This method was tested by reproducing and expanding upon published respiratory studies. The results showed that a user could, with little lung function knowledge, effectively discover new knowledge from a large dataset with the incorporation of complex existing knowledge. Each data analysis study was also performed efficiently, finding results rapidly. The discovered knowledge was reinforced by recent literature, and some of the analyses seem to present relevant findings for current research, despite the relative inexperience of the user. A notable advantage of the method is that it also incorporates newly discovered knowledge automatically, allowing progressive knowledge discovery.

While the method is not considered as rigorous as more specifically designed studies, the use of retrospective data analysis is widely recognised in the field, and the results suggest that the presented approach provides an efficient and effective way to perform this type of analysis. Certain restrictions on the data mean that some findings made from the discovered knowledge are not considered to be conclusive; but the results still suggest that the method can be effectively used as an exploratory data analysis tool, testing and expanding upon research hypotheses, with potential for the discovery of new relationships to assist in the development of the initial idea. It is also expected that the method can improve through the addition of more expansive datasets and further data analysis functions, which should improve the results and increase the scope of application.

Finally, Chapter 5 presented a method for quantifiably comparing the knowledge of multiple experts. Comparing the results of applying each expert's knowledge to the dataset allowed the identification of conflicts, the magnitude of each and the details causing each conflict, as well as the information needed to resolve them. The results showed that the method could successfully identify and quantify the differences between the experts' acquired knowledge, and provide the information needed to

resolve them, such that a consolidated knowledge base was developed. The same method was also applied to the acquired knowledge of a group of medical students, as a knowledge comparison tool for improving, and identifying weaknesses in, student knowledge. The students described beneficial learning outcomes through the acquisition process, as it provided them an avenue to apply and develop their own practical understanding of previously learned theory, and highlighted inconsistencies in their asserted explanations. The quantified comparisons likewise showed potential for increasing the students' understanding by discovering exactly where their knowledge was lacking. The method could also provide examples of cases where the students' understanding would result in incorrect interpretations, both demonstrating the flaws and, once corrected, allowing the students to identify for themselves the relevant patterns in real cases, rather than simply memorising a rule. The comparison method also showed some promise as an assessment tool, through the calculation of general accuracy measures and magnitudes for each difference in knowledge, defined by their results for a set of real cases. It was established from the expert knowledge comparisons that the magnitude of differences in knowledge were difficult to estimate by looking at rule conditions alone. Although successful as a learning tool, and showing promise as an assessment tool, further work is required to determine a reliable means of application for assessment, given the multi-faceted comparison results.

6.1 Further Work

Many aspects of the work presented in this thesis require further evaluation, particularly in comparison to other approaches. The knowledge acquisition results leave many unanswered questions about the effects of the implementation changes, and how best to overcome difficulties faced. How to maintain the case-based approach to knowledge acquisition, while allowing the expert some freedom to define rules without a case, is an interesting problem. As the effects are not disastrously detrimental, the author suspects that the resolution to this problem will be in balancing and minimising the impact of rule-based acquisition, rather than removing it entirely.

How to maximise the impact of the statistical validation is also an area to be explored. The positive results from the student knowledge acquisition shows that

this method has some promise, but the unwillingness of most participants to investigate the statistics highlights that there is further work to be done.

The statistics derived from the dataset, both for validation and for knowledge discovery, can likely be improved. The statistics used in this study only provide an example of the types of calculations that can be performed. Similarly, the data mining techniques used (basic association rule and information gain calculations) provide only an example. Many different data mining techniques could be employed to assist the user further.

Perhaps most promising for future research and development is the application of the knowledge comparison method as a learning and assessment tool. Testing the method with students found that most participants felt they had benefited from the process, particularly in gaining practical experience examining cases. Notably, the students described that they had learned from the process before they had seen any comparison between their knowledge and the experts'; in fact, before they had seen any expert knowledge. It would be expected that by incorporating these comparisons the learning outcomes could be improved significantly. The potential for the method to automatically identify significant (or subtle) problems with a student's knowledge, and then automatically identify pertinent cases to provide the student with, is an area that shows much potential for further work.

6.2 Conclusion

Although presented as individual experiments in separate chapters, the methods described in this thesis are, in fact, components of a single system. This system allows experts to compare, consolidate, and develop their knowledge by intuitively interpreting data, both at an individual case level and by examining wider data trends. The knowledge comparison results showed that knowledge can be tested against a more experienced expert, or experts of a similar level of experience can contrast their differing approaches to data interpretation; in both situations differences are identified and quantified, and possible solutions can be explored with evidence; thus allowing collaborative knowledge acquisition which assists in identifying and resolving conflicts, and improving each expert's knowledge. The student participants indicated that even with minimal use of the system, without any focus on being taught, they learned from the experience. This provides very

promising evidence that this system helps to develop knowledge and could be applied as a learning tool.

In the same framework, experts can easily test hypotheses against the dataset; the data mining tools will assist in identifying interesting relationships within the data, based on the knowledge that the expert has described. As more definitions are added to the knowledge base, it becomes easier to test specific relationships, and automatically identify interesting or unexpected relationships between data groups. Any new relationships that are discovered are automatically included in the knowledge base, allowing them to be applied immediately to either discover further knowledge, or as a benefit to an expert system. The successful discovery of new knowledge in the lung function domain shows that the method can effectively acquire complex knowledge and apply it to a knowledge discovery task. That this was performed by a novice in the domain provides more evidence that it is the acquired knowledge which allowed the discovery of useful results.

References

Aamodt, A., & Plaza, E. (1994). Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Communications, 7*(1), 33-59.

Abe, H., & Yamaguchi, T. (2005). *Implementing an Integrated Time-Series Data Mining Environment - A Case Study of Medical KDD on Chronic Hepatitis.* Paper presented at the First International Conference on Complex Medical Engineering (CME2005), Takamatsu, Kagawa, Japan.

Adejuwon, A., & Mosavi, A. (2010). Domain Driven Data Mining–Application to Business. *International Journal of Computer Science Issues, 7*(4).

Agahi, A. (2007). *Patterns of Lung Function in Health and Disease.* Honours Thesis, University of Tasmania, Hobart.

Aggarwal, A., Gupta, D., Behera, D., & Jindal, S. (2006). Comparison of fixed percentage method and lower confidence limits for defining limits of normality for interpretation of spirometry. *Respiratory Care, 51*(7), 737.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases *Proceedings of the 1993 ACM SIGMOD international conference on Management of data* (pp. 207-216). Washington, D.C., United States: ACM Press.

Aikins, J., Kunz, J., Shortliffe, E., & Fallat, R. (1983). PUFF: an expert system for interpretation of pulmonary function data. *Computers and Biomedical Research, 16*(3), 199-208.

American Thoracic Society. (1991). Lung function testing: selection of reference values and interpretative strategies. *American Review of Respiratory Diseases, 144*, 1202-1218.

Anderson, T. (2004). Teaching in an Online Learning Context. *Theory and Practice of Online Learning*, 273-294.

Anees, S., Coyle, K., & Aldrich, T. (2009). Measuring Total Lung Capacity (TLC) as Single-Breath Alveolar Volume (VAsb), with Correction for Maldistributed Ventilation Using Maximum Mid-Expiratory Flow Rate (MMEFR). *American Journal of Respiratory and Critical Care Medicine, 179*(1 MeetingAbstracts), A4419.

Anthonisen, N., & Wright, E. (1986). Bronchodilator response in chronic obstructive pulmonary disease. *The American review of respiratory disease, 133*(5), 814.

Arshadi, N., & Jurisica, I. (2005). Data mining for case-based reasoning in high-dimensional biological domains. *IEEE Transactions on Knowledge and Data Engineering*, 1127-1137.

Bachant, J., & McDermott, J. (1984). RI Revisited: Four Years in the Trenches. *AI Magazine, 5*(3), 21-32.

Barker, V., O'Connor, D., Bachant, J., & Soloway, E. (1989). Expert systems for configuration at Digital: XCON and beyond. *Communications of the ACM, 32*(3), 298-318.

Barletta, R., & Mark, W. (1988). *Explanation-based indexing of cases.* Paper presented at the Case-Based Reasoning Workshop, Palo Alto.

Barricelli, N. (1957). Symbiogenetic evolution processes realized by artificial methods. *Methodos, 9*(35-36), 143-182.

Bayardo Jr., R. J., & Agrawal, R. (1999). Mining the most interesting rules *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 145-154). San Diego, California, United States: ACM Press.

Ben Saad, H., Préfaut, C., Tabka, Z., Zbidi, A., & Hayot, M. (2008). The forgotten message from gold: FVC is a primary clinical outcome measure of bronchodilator reversibility in COPD. *Pulmonary pharmacology & therapeutics, 21*(5), 767-773.

Bentley, P., & Corne, D. (2002). *Creative evolutionary systems*. San Francisco: Morgan Kaufmann Pub.

Berkhin, P. (2006). A survey of clustering data mining techniques. *Grouping Multidimensional Data*, 25-71.

Beydoun, G., & Hoffmann, A. (2000). Incremental acquisition of search knowledge. *International Journal of Human-Computer Studies, 52*(3), 493-530.

Beydoun, G., Hoffmann, A., Breis, J. T. F., Bejar, R. M., Valencia-Garcia, R., & Aurum, A. (2005). Cooperative modelling evaluated. *International Journal of Cooperative Information Systems, 14*(1), 45-71.

Bindoff, I. K. (2010). *Multiple Classification Ripple Round Rules: Classifications as Conditions.* PhD Thesis, University of Tasmania, Hobart.

Bleecker, E. R. (2004). Similarities and Differences in Asthma and COPD: the Dutch hypothesis. *Chest, 126*(2 suppl 1), 93S.

Bobrow, D. G., Sanjay, M., & Stefik, M. J. (1986). Expert Systems: Perils and Promises. *Communications of the ACM, 29*(9), 880-894.

Borg, B. M., Reid, D. W., Walters, E. H., & Johns, D. P. (2004). Bronchodilator reversibility testing: laboratory practices in Australia and New Zealand. *Medical journal of Australia, 180*(12), 610-613.

Brachman, R., & Anand, T. (1996). *The Process of Knowledge Discovery in Databases: A Human-Centered Approach.* Paper presented at the Advances in Knowledge Discovery and Data Mining, California.

Brin, S., Motwani, R., Ullman, J., & Tsur, S. (1997). *Dynamic itemset counting and implication rules for market basket data.* Paper presented at the 1997 International Conference on Management of Data.

Brown, A. L., & Palincsar, A. S. (1989). Guided, Cooperative Learning and Individual Knowledge Acquisition. *Knowing, learning, and instruction: essays in honor of Robert Glaser*, 393-451.

Buchanan, B., Mitchell, T., & SCIENCE., S. U. C. D. O. C. (1977). *Model-directed learning of production rules*: Computer Science Department, Stanford University.

Buchanan, B. G. (1986). Expert systems: working systems and the research literature. *Expert Systems, 3*(1), 32-50.

Buchanan, B. G., Barstow, D., Bechtal, R., Bennett, J., Clancey, W., Kulikowski, C., et al. (1983). Constructing an expert system. *Building Expert Systems*, 127-167.

Buchanan, B. G., & Feigenbaum, E. A. (1978). DENDRAL and Meta-DENDRAL: Their Applications Dimension. *Artificial Intelligence, 11*(1978), 5-24.

Buchanan, B. G., & Shortliffe, E. H. (1984). *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project*. Boston: Addison-Wesley.

Buchanan, B. G., & Sutherland, G. (1968). *Heuristic DENDRAL: a program for generating explanatory hypotheses in organic chemistry*. New York: Stanford University California Department of Computer Science.

Bulkeley, W. (1990). Technology: Bright outlook for artificial intelligence yields to slow growth and big cutbacks. *Wall Street Journal, Section B*, 1.

Burns, C., & Scheinhorn, D. (1984). Evaluation of single-breath helium dilution total lung capacity in obstructive lung disease. *The American review of respiratory disease, 130*(4), 580.

Burrows, B., Bloom, J. W., Traver, G. A., & Cline, M. G. (1987). The course and prognosis of different forms of chronic airways obstruction in a sample from the general population. *New England Journal of Medicine, 317*(21), 1309-1314.

Caballero, B. (2007). The global epidemic of obesity: an overview. *Epidemiologic reviews, 29*(1), 1.

Carbonell, J., Michalski, R., & Mitchell, T. (1983). An overview of machine learning. *Machine learning: An artificial intelligence approach, 1*, 3–23.

Cassam, Q. (2009). What is knowledge? *Royal Institute of Philosophy Supplements, 84*(64), 101-120.

Chein, M., & Mugnier, M. L. (2008). *Graph-based knowledge representation: computational foundations of conceptual graphs*. London: Springer-Verlag New York Inc.

Chi, R. H., & Kiang, M. Y. (1991). An integrated approach of rule-based and case-based reasoning for decision support *Proceedings of the 19th annual conference on Computer Science* (pp. 255-267). San Antonio, Texas, United States: ACM Press.

Cios, K., & Kacprzyk, J. (2001). *Medical data mining and knowledge discovery*. Denver: Physica-Verlag.

Cios, K., & Moore, G. W. (2002a). Uniqueness of medical data mining. *Artificial Intelligence in Medicine, 26*(1-2), 1-24.

Cios, K., & Moore, W. (2002b). Uniqueness of medical data mining. *Artificial Intelligence in Medicine, 26*(1-2), 1-24.

Clancey, W. J. (1984). Knowledge acquisition for classification expert systems *Proceedings of the 1984 annual conference of the ACM on The fifth generation challenge* (pp. 11-14): ACM Press.

Clancey, W. J. (1993). Situated action: A neuropsychological interpretation (Response to Vera and Simon). *Cognitive Science, 17*(1), 87-116.

Collen, J., Greenburg, D., Holley, A., King, C., & Hnatiuk, O. (2008). Discordance in spirometric interpretations using three commonly used reference equations vs National Health and Nutrition Examination Study III. *Chest, 134*(5), 1009.

Compton, P. (1992). *Insight and knowledge.* Paper presented at the AAAI Spring Symposium: Cognitive Aspects of Knowledge Acquisition, Stanford University.

Compton, P., & Edwards, G. (1994). *A 2000 Rule Expert System Without a Knowledge Engineer.* Paper presented at the Proceedings of the 8th AAAI-Sponsored Ban Knowledge Acquisition for Knowledge-Based Systems Workshop, Banff, Canada.

Compton, P., Edwards, G., Kang, B., Lazarus, L., Malor, R., Menzies, T., et al. (1991). *Ripple down rules: possibilities and limitations.* Paper presented at

the The 6th Knowledge Acquisition for Knowledge-Based Systems Workshop, Canada.

Compton, P., Edwards, G., Kang, B., Lazarus, L., Malor, R., Preston, P., et al. (1992). Ripple down rules: Turning knowledge acquisition into knowledge maintenance. *Artificial Intelligence in Medicine, 4*(6), 463-475.

Compton, P., & Jansen, R. (1989). *A philosophical basis for knowledge acquisition.* Paper presented at the European Knowledge Acquisition for Knowledge-Based Systems, Paris.

Compton, P., Kang, B., Preston, P., & Mulholland, M. (1993). *Knowledge Acquisition without Analysis.* Paper presented at the Knowledge Acquisition for Knowledge-Based Systems, Springer Verlag.

Compton, P., Peters, L., Edwards, G., & Lavers, T. (2006). Experience with ripple-down rules. *Knowledge-Based Systems, 19*(5), 356-362.

Compton, P., Preston, P., Edwards, G., & Kang, B. (1996). *Knowledge based systems that have some idea of their limits.* Paper presented at the 10th Knowledge Acquisition for Knowledge-Based Systems Workshop, Canada.

Compton, P., Ramadan, Z., Preston, P., Le-Gia, T., Chellen, V., & Mullholland, M. (1998). *A trade-off between domain knowledge and problem solving method power.* Paper presented at the 11th Knowledge Acquisition for Knowledge-Based Systems Workshop.

Cotes, J., & Leathart, G. (1993). *Lung function: assessment and application in medicine*. Oxford: Wiley-Blackwell.

Craik, F. I. M., & Tulving, E. (2004). Depth of processing and the retention of words in episodic memory. *Cognitive psychology: key readings, 296*.

Crapo, R., Gardner, R., & Clausen, J. (1987). Single breath carbon monoxide diffusing capacity (transfer factor): recommendations for a standard technique. *American Review of Respiratory Diseases, 136*, 1299-1307.

Crapo, R., & Morris, A. (1981). Standardized single breath normal values for carbon monoxide diffusing capacity. *American Review of Respiratory Diseases, 123*(2), 185.

Crapo, R., Morris, A., & Gardner, R. (1981). Reference spirometric values using techniques and equipment that meet ATS recommendations. *American Review of Respiratory Diseases, 123*(6), 659.

Creighton, C., & Hanash, S. (2003). Mining gene expression databases for association rules. *Bioinformatics, 19*(1), 79.

Culver, B. (2006). Interpretation of spirometry: we can do better than the GOLD standard. *Respiratory care, 51*(7), 719.

Dai, J. Y., Yang, D. L., Wu, J., & Hung, M. C. (2008). An Efficient Data Mining Approach on Compressed Transactions. *World Academy of Science, Engineering and Technology, April*(40), 522-529.

Davies, B., & Darbyshire, I. (1984). The use of expert systems in process-planning. *CIRP Annals-Manufacturing Technology, 33*(1), 303-306.

Davis, R., Buchanan, B., & Shortliffe, E. H. (1977). Production rules as a representation for a knowledge-based consultation program. *Artificial Intelligence, 8*(1), 15-45.

Davison, B., & Hirsh, H. (1998). *Predicting Sequences of User Actions.* Paper presented at the Predicting the Future: AI Approaches to Time-Series Analysis.

Dazeley, R., & Kang, B. (2003). *Weighted MCRDR: Deriving Information about Relationships between Classifications in MCRDR.* Paper presented at the

16th Australian Joint Conference on Artificial Intelligence (AI'03), Perth, Australia.

Dazeley, R., & Kang, B. (2004). *An Online Classification and Prediction Hybrid System for Knowledge Discovery in Databases.* Paper presented at the The 2nd International Conference on Artificial Intelligence in Science and Technology, Hobart, Australia.

Dieng, R. (1997). Comparison of conceptual graphs for modelling knowledge of multiple experts: application to traffic accident analysis. *Rapport de Recherche-Iintitut National de Recherche en Iinformatique et en Automatique, 3161*.

Duda, R., Hart, P., & Nilsson, N. (1976). *Subjective Bayesian methods for rule-based inference systems.* Paper presented at the 1976 National Computer Conference, New York.

Duda, R., & Shortliffe, E. (1983). Expert systems research. *Science, 220*(4594), 261-268.

Duffy, T. M., & Cunningham, D. J. (1996). Constructivism: Implications for the design and delivery of instruction. *Handbook of research for educational communications and technology, 171*.

Edwards, G., Compton, P., Malor, R., Srinivasan, A., & Lazarus, L. (1993). PEIRS: a pathologist-maintained expert system for the interpretation of chemical pathology reports. *Pathology*(25), 27-34.

Eliasson, O., & Degraff Jr, A. (1985). The use of criteria for reversibility and obstruction to define patient groups for bronchodilator trials. Influence of clinical diagnosis, spirometric, and anthropometric variables. *The American review of respiratory disease, 132*(4), 858.

Escovar, E. L. G., Yaguinuma, C. A., & Biajiz, M. (2006). Using Fuzzy Ontologies to Extend Semantically Similar Data Mining. *Proceedings of the XXI Simpósio Brasileiro de Banco de Dados (SBBD 2006)*, 16–30.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). *Knowledge discovery and data mining: Towards a unifying framework.* Paper presented at the Second International Conference on Knowledge Discovery and Data Mining Portland, Oregon.

Féret, M., & Glasgow, J. (1997). Combining Case-Based and Model-Based Reasoning for the Diagnosis of Complex Devices. *Applied Intelligence, 7*(1), 57-78.

Ferguson, G., Enright, P., Buist, A., & Higgins, M. (2000). Office Spirometry for Lung Health Assessment in Adults: A Consensus Statement From the National Lung Health Education Program. *Chest*(117), 1146-1161.

Ferris, B. (1978). Epidemiology standardization project: recommended standardized procedures for pulmonary function testing. *American Review of Respiratory Diseases, 118*(6), 57-59.

Fix, E., & Hodges, J. (1951). *Disciminatory analysis---nonparametric discrimination; consistency properties*. Randolph Field, Texas: USAF School of Aviation Medicine.

Fox, R., & Bennett, N. (1998). Continuing medical education: learning and change: implications for continuing medical education. *British Medical Journal, 316*(7129), 466.

Fransella, F., Bell, R., & Bannister, D. (1979). *A Manual for Repertory Grid Technique* (2 ed.). London: Academic Press.

Frawley, W., Piatetsky-Shapiro, G., & Matheus, C. (1992). Knowledge Discovery in Databases: An Overview. *AI Magazine, 13*(3), 57-70.

Freitag, D. (2000). Machine learning for information extraction in informal domains. *Machine Learning, 39*(2), 169-202.

Freitas, A. A. (1999). On rule interestingness measures. *Knowledge-Based Systems, 12*(5-6), 309-315.

Friedberg, R. (1958). A learning machine: Part I. *IBM Journal of Research and Development, 2*(1), 2-13.

Friedman, G. (1959). Digital simulation of an evolutionary process. *General Systems Yearbook, 4*, 171-184.

Fritzke, B. (1993). *Kohonen feature maps and growing cell structures-a performance comparison.* Paper presented at the Advances in Neural Information Processing Systems 5, San Mateo, California.

Fukuda, T., Morimoto, Y., Morishita, S., & Tokuyama, T. (1996). *Data mining using two-dimensional optimized association rules: Scheme, algorithms, and visualization.* Paper presented at the ACM SIGMOD Conference on Management of Data.

Gaines, B., & Boose, J. (1988). *Knowledge Acquisition for Knowledge-Based Systems* (Vol. 1). Orlando, FL, USA: Academic Press, Inc.

Gaines, B., & Compton, P. (1992). *Induction of Ripple Down Rules.* Paper presented at the Proceedings of the 5th Australian Conference on Artical Intelligence, Hobart, Australia.

Gaines, B. R. (1987). *Rapid prototyping for expert systems.* Paper presented at the Proceedings from First International Conference on Expert Systems and the Leading Edge in Production Planning and Control., Menlo Park, California.

Gardner, M., & Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences. *Atmospheric environment, 32*(14-15), 2627-2636.

Gialamas, A., Beilby, J., Pratt, N., Henning, R., Marley, J., & Roddick, J. (2003). Investigating tiredness in Australian general practice-Do pathology tests help in diagnosis? *Australian family physician, 32*(8), 663-666.

Gill, T. G. (1995). Early Expert Systems: Where Are They Now? *MIS Quarterly, 19*(1), 51-81.

Goebel, M., & Gruenwald, L. (1999). A survey of data mining and knowledge discovery software tools. *SIGKDD Explorations, 1*(1), 20-33.

GOLD. (2008). *Global Strategy for the Diagnosis, Management and Prevention of COPD*: Global Initiative for Chronic Obstructive Lung Disease.

Golding, A., & Rosenbloom, P. (1996). Improving accuracy by combining rule-based and case-based reasoning. *Artificial Intelligence, 87*(1), 215-254.

Goldman, H., & Becklake, M. (1959). Respiratory function tests; normal values at median altitudes and the prediction of normal results. *American review of tuberculosis, 79*(4), 457.

Goldman, J., Chu, W., Parker, D., & Goldman, R. (2008). Term domain distribution analysis: a data mining tool for text databases. *Nuklearmedizin, 47*(1), 48-55.

Good, T. L., & Brophy, J. E. (1990). *Educational psychology. A realistic approach* (4 ed.). New York: Addison Wesley.

Grefenstette, J., Ramsey, C., & Schultz, A. (1990). Learning sequential decision rules using simulation models and competition. *Machine Learning, 5*(4), 355-381.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*, 199-220.

Hall, M., & Smith, L. (1998). *Practical feature subset selection for machine learning.* Paper presented at the 21st Australian Computer Science Conference, Perth, Australia.

Hand, D., & Vinciotti, V. (2003). Choosing k for two-class nearest neighbour classifiers with unbalanced classes. *Pattern Recognition Letters, 24*(9-10), 1555-1562.

Hankinson, J., Odencrantz, J., & Fedan, K. (1999). Spirometric reference values from a sample of the general US population. *American journal of respiratory and critical care medicine, 159*(1), 179.

Hardie, J., Buist, A., Vollmer, W., Ellingsen, I., Bakke, P., & Mørkve, O. (2002). Risk of over-diagnosis of COPD in asymptomatic elderly never-smokers. *European Respiratory Journal, 20*(5), 1117.

Hart, P., & Center, A. (1977). PROSPECTOR--A Computer-Based Consultation System for Mineral Exploration. *Mathematical Geology, 10*(5), 589-610.

Hartigan, J. A. (1975). *Clustering algorithms*. New York: John Wiley & Sons.

Hayward, S. (1985). Is a decision tree an expert system? *Research and Development in Expert Systems*, 185-192.

Heckerman, D. (1995). A Tutorial on Learning With Bayesian Networks. *Learning in Graphical Models*, 301-354.

Hidber, C. (1999). *Online Association Rule Mining.* Paper presented at the ACM SIGMOD Conference on Management of Data.

Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. from American Association for the Advancement of Science: http://www.sciencemag.org/content/early/2011/02/09/science.1200970

Holland, J. (1975). *Adaptation in Natural and Artificial Systems*. Michigan: University Michigan Press, Ann Arbor.

Hong, T.-P., Wang, T.-T., Wang, S.-L., & Chien, B.-C. (2000). Learning a coverage set of maximally general fuzzy rules by rough sets. *Expert Systems with Applications, 19*(2), 97-103.

Horn, K., Compton, P., Lazarus, L., & Quinlan, J. (1985). An expert system for the interpretation of thyroid assays in a clinical laboratory. *Australian computer journal, 17*(1), 7-11.

Horvitz, E. (1986). *Toward a science of expert systems.* Paper presented at the 18th Symposium on the Interface of Computer Science and Statistics, Ft Collins, Colorado.

Huang, G., Saratchandran, P., & Sundararajan, N. (2005). A generalized growing and pruning RBF (GGAP-RBF) neural network for function approximation. *Neural Networks, IEEE Transactions on, 16*(1), 57-67.

Hughes, D., & Empey, D. (1981). *Lung Function for the Clinician*. London: Academic Press Grune and Stratton.

Hunt, E., Marin, J., & Stone, P. (1966). *Experiments in Induction*. New York: Academic Press.

Jain, A., Murty, M., & Flynn, P. (1999). Data Clustering: a review. *ACM computing surveys (CSUR), 31*(3), 264-323.

Jansen, B., & Compton, P. (1988). *The knowledge dictionary: A relational tool for the maintenance of expert systems.* Paper presented at the 5th Generation Computer Systems, Tokyo, Japan.

Jenkins, C., & Young, I. (2004). Assessing bronchodilator reversibility: agreed standards are urgently needed. *The Medical journal of Australia, 180*(12), 605.

Johns, M. V. (1961). An empirical Bayes approach to non-parametric two-way classification. *Studies in Item Analysis and Prediction*, 221-232.

Jones, R. L., & Nzekwu, M. M. U. (2006). The Effects of Body Mass Index on Lung Volumes. *Chest, 130*(3), 827-833.

Kang, B. (1996). *Validating Knowledge Acquisition: Multiple Classification Ripple Down Rules.* PhD Thesis, University of New South Wales, Sydney.

Kang, B., & Compton, P. (1992). *Knowledge Acquisition in Context: the Multiple Classification Problem.* Paper presented at the Proceedings of the Pacific Rim International Conference on Artificial Intelligence, Seoul.

Kang, B., Compton, P., & Preston, P. (1995, Feb 26 - March 3). *Multiple Classification Ripple Down Rules: Evaluation and Possibilities.* Paper presented at the Proceedings 9th Banff Knowledge Acquisition for Knowledge Based Systems Workshop, Banff.

Kent, J. (1983). Information gain and a general measure of correlation. *Biometrika, 70*(1), 163.

Kerstjens, H. (2004). The GOLD classification has not advanced understanding of COPD. *American journal of respiratory and critical care medicine, 170*(3), 212.

Khinchin, A. I. (1957). *Mathematical foundations of information theory*. New York: Dover Publications.

Knudson, R., Slatin, R., Lebowitz, M., & Burrows, B. (1976). The maximal expiratory flow-volume curve. Normal standards, variability, and effects of age. *American Review of Respiratory Diseases, 113*(5), 587.

Kodaz, H., Özsen, S., Arslan, A., & Günes, S. (2009). Medical application of information gain based artificial immune recognition system (AIRS): Diagnosis of thyroid disease. *Expert Systems with Applications, 36*(2), 3086-3092.

Kolodner, J. (1991). Improving Human Decision Making through Case-Based Decision Aiding. *AI Magazine, 12*(2), 52-68.

Kolodner, J. (1992). An introduction to case-based reasoning. *Artificial Intelligence Review, 6*(1), 3-34.

Kolodner, J., Simpson, R., & Sycara-Cyranski, K. (1985). *A process model of cased-based reasoning in problem solving.* Paper presented at the 9th International Joint Conference on Artificial Intelligence, Los Angeles, California.

Kononenko, I., Bratko, I., & Roskar, E. (1984). *Experiments in automatic learning of medical diagnostic rules.* Paper presented at the International School for the Synthesis of Expert's Knowledge Workshop, Bled, Slovenia.

Kotsifakos, E. E., Marketos, G., & Theodoridis, Y. (2008). A framework for integrating ontologies and pattern-bases. In H. O. Nigro, S. G. Cisaro & D. Xodo (Eds.), *Data Mining with Ontologies: Implementations, Findings, and Frameworks. Information Science Reference*: Idea Group Inc., Hershey.

Kowalski, A. (1991). Case-based reasoning and the deep structure approach to knowledge representation *Proceedings of the 3rd international conference on Artificial intelligence and law* (pp. 21-30). Oxford, England: ACM Press.

Kullback, S., & Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*, 79-86.

Kurgan, L. A., & Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review, 21*(01), 1-24.

Kurniawati, R., Jin, J., & Shepherd, J. (1998). Efficient nearest-neighbour searches using weighted euclidean metrics. *Proceedings of the 16th British National Conference on Databases: Advances in Databases*, 64-76.

Kusiak, A., Kern, J., Kernstine, K., & Tseng, B. (2002). Autonomous decision-making: A data mining approach. *Information Technology in Biomedicine, IEEE Transactions on, 4*(4), 274-284.

Laszlo, G. (1994). *Pulmonary Function: A Guide for Clinicians*. Cambridge: Cambridge University Press.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. New York: Cambridge University Press.

Lavra , N., Flach, P., & Zupan, B. (1999). *Rule evaluation measures: A unifying view.* Paper presented at the 9th International Workshop on Inductive Logic Programming, Heidelberg.

Lederberg, J., Feigenbaum, E., & CALIF., S. U. (1967). *Mechanization of inductive inference in organic chemistry*: Defense Technical Information Center.

Lenat, D., Prakash, M., & Shepherd, M. (1985). CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *AI Magazine, 6*(4), 65.

Lenca, P., Vaillant, B., & Lallich, S. (2006). *On the robustness of association rules.* Paper presented at the 2nd IEEE International Conference on Cybernetics and Intelligent Systems and Robotics, Automation and Mechatronics, Bangkok, Thailand.

Ling, T. R. (2006). *An Incremental Learning Method for Data Mining from Large Databases*. Honours Thesis, University of Tasmania, Hobart.

Liou, Y. I. (1990). *Knowledge acquisition: issues, techniques, and methodology.* Paper presented at the Conference on Trends and Directions in Expert Systems, Orlando, Florida, United States.

Liu, B., Hsu, W., & Chen, S. (1997). *Using general impressions to analyze discovered classification rules.* Paper presented at the 3rd International Conference on Knowledge Discovery and Data Mining, Newport Beach, California.

Luconi, F., Malone, T., Morton, S., & Michael, S. (1984). *Expert systems and expert support systems: the next challenge for management*. Cambridge, Massachusetts: Center for Information Systems Research, Sloan School of Management, Massachusetts Institute of Technology.

MacKay, D. (1992). Information-based objective functions for active data selection. *Neural computation, 4*(4), 590-604.

MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations.* Paper presented at the Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, California, USA.

Manago, M., Althoff, K., Auriol, E., Traphoner, R., Stefan, W., Conruyt, N., et al. (1993). *Induction and Reasoning from Cases.* Paper presented at the First European Workshop on Case-Based Reasoning, Kaiserslautern, Germany.

Marinica, C., & Guillet, F. (2009). *Improving Post-Mining of Association Rules with Ontologies.* Paper presented at the 13th International Conference on Applied Stochastic Models and Data Analysis, Vilnius, Lithuania.

Marinica, C., Guillet, F., & Briand, H. (2008). *Post-processing of discovered association rules using ontologies*.

Masić, I., Ridanović, Z., & Pandza, H. (1995). Medical expert systems. *Medicinski arhiv, 49*(3-4), 107.

Matheus, C. J., Chan, P. K., & Piatetsky-Shapiro, G. (2002). Systems for knowledge discovery in databases. *Knowledge and Data Engineering, IEEE Transactions on, 5*(6), 903-913.

McCarthy, J., Hayes, P., & SCIENCE., S. U. C. D. O. C. (1968). *Some philosophical problems from the standpoint of artificial intelligence*: Stanford University.

McGarry, K. (2005). A survey of interestingness measures for knowledge discovery. *The Knowledge Engineering Review, 20*(01), 39-61.

MedGraphics. (2011). MedGraphics Pulmonary Consult™ Software. Retrieved 20th January, 2011, from http://www.medgraphics.com/datasheet_pconsult.html

Meneely, G., Renzetti, A., Steele, J., Wyatt, J., & Harris, H. (1962). Chronic bronchitis, asthma, and pulmonary emphysema. A statement by the committee on diagnostic standards for non-tuberculous respiratory diseases. *American Review of Respiratory Diseases, 85*, 762-768.

Mezirow, J. (1991). *Transformative Dimensions of Adult Learning*. San Francisco: Jossey-Bass.

Michalski, R. (1978). *Pattern recognition as knowledge-guided computer induction*: Dept. of Computer Science, University of Illinois at Urbana-Champaign.

Miller, A. (1987). *Pulmonary Function Tests*. Orlando: Grune and Stratton.

Miller, M., Crapo, R., Hankinson, J., Brusasco, V., Burgos, F., Casaburi, R., et al. (2005). General considerations for lung function testing. *European Respiratory Journal, 26*(1), 153.

Minsky, M. (1961). Steps Toward Artificial Intelligence. *Proceedings of the Institute of Radio Engineers, 49*(1), 8-30.

Minsky, M., & Papert, S. (1988). *Perceptrons*: MIT press.

Misra, A., Sowmya, A., & Compton, P. (2011). Incremental system engineering using process networks. *Knowledge Management and Acquisition for Smart Systems and Services*, 150-164.

Mitchell, M. (1998). *An introduction to genetic algorithms*. Cambridge: The MIT press.

Mitchell, M., & Renzetti Jr, A. (1968). Evaluation of a single-breath method of measuring total lung capacity. *American Review of Respiratory Diseases, 97*(4), 571.

Mitchell, T. (1997). Artificial Neural Networks. In T. Mitchell (Ed.), *Machine Learning* (pp. 82-112): McGraw-Hill.

Mulholland, M., Preston, P., Sammut, C., Hibbert, B., & Compton, P. (1993). *An Expert System for Ion Chromatography developed using Machine Learning and Knowledge in Context.* Paper presented at the 6th International Conference on Industrial and Engineering Applications of Artifical Intelligence and Expert Systems, Edinburgh.

Musen, M. A., Shahar, Y., & Shortliffe, E. H. (2006). Clinical decision-support systems. *Biomedical Informatics*, 698-736.

NACA. (2005). How to Perform Spirometry. Retrieved 22nd January, 2011, from http://www.nationalasthma.org.au/content/view/331/418/

Nilsson, N. J. (1965). *Learning machines*. New York: McGraw-Hill.

O'Donnell, D. E., Deesomchok, A., Lam, Y. M., Guenette, J. A., Amornputtisathaporn, N., Forkert, L., et al. (2011). Effects of Body Mass Index on Static Lung Volumes in Patients with Airway Obstruction. from American College of Chest Physicians:

Ohsaki, M., Sato, Y., Kitaguchi, S., Yokoi, H., & Yamaguchi, T. (2004). Comparison between objective interestingness measures and real human interest in medical data mining. *Innovations in Applied Artificial Intelligence*, 1072-1081.

Ordonez, C. (2006). Association rule discovery with the train and test approach for heart disease prediction. *Information Technology in Biomedicine, IEEE Transactions on, 10*(2), 334-343.

Orriols-Puig, A., Casillas, J., & Bernadó-Mansilla, E. (2008). Genetic-based machine learning systems are competitive for pattern recognition. *Evolutionary Intelligence, 1*(3), 209-232.

Ortega y Gasset, J., & García-Gómez, J. (2002). *What is knowledge?* New York: State University of New York Press.

Patterson, A., & Niblett, T. (1982). ACLS user manual. Glasgow, Scotland: Intelligent Terminal Ltd.

Pears, D. (1971). *What is knowledge?* : Harper & Row.

Pecora, L., Bernstein, I., & Feldman, D. (1968). Comparison of the components of diffusing capacity utilizing the effective alveolar volume in patients with emphysema and chronic asthma. *The American Journal of the Medical Sciences, 256*(2), 69.

Pellegrino, R., Viegi, G., Brusasco, V., Crapo, R., Burgos, F., Casaburi, R., et al. (2005). Interpretative strategies for lung function tests. *European Respiratory Journal, 26*(5), 948.

Pernin, C. G. (2008). *Allocation of Forces, Fires, and Effects Using Genetic Algorithms*. Santa Monica, California: Rand Arroyo Center.

Piaget, J. (1972). *Psychology and epistemology: Towards a theory of knowledge*. London: Allen Lane.

Piatetsky-Shapiro, G. (1990). Knowledge discovery in real databases: A report on the IJCAI-89 Workshop. *AI Magazine, 11*(4), 68.

Piatetsky-Shapiro, G. (1991). Discovery, Analysis, and Presentation of Strong Rules. *Knowledge Discovery in Databases*, 229-238.

Piatetsky-Shapiro, G. (2000). Knowledge Discovery in Databases: 10 years after. *ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) Explorations, 1*(2), 59-61.

Piatetsky-Shapiro, G., Matheus, C., Smyth, P., & Uthurusamy, R. (1994). Kdd-93: Progress and challenges in knowledge discovery in databases. *AI Magazine, 15*(3), 77.

Piatetsky-Shapiro, G., & Matheus, C. J. (1994). *The Interestingness of Deviations.* Paper presented at the AAAI-94 Workshop on Knowledge Discovery in Databases, Seattle, Washington.

Pierce, R. J., Hillman, D., Young, I. H., O'donoghue, F., Zimmerman, P. V., West, S., et al. (2005). Respiratory function tests and their application. *Respirology, 10*(s2), S1-S19.

Pohle, C. (2003). *Integrating and updating domain knowledge with data mining.* Paper presented at the Very Large Database (VLDB) PhD Workshop, Berlin, Germany.

Prather, J., Lobach, D., Goodwin, L., Hales, J., Hage, M., & Hammond, W. (1997). *Medical data mining: knowledge discovery in a clinical data warehouse.* Paper presented at the 1997 Annual Conference of the American Medical Informatics Association, Philadelphia.

Prerau, D. (1985). Selection of an appropriate domain for an expert system. *AI Magazine, 6*(2), 26.

Pribor, H. (1989). Expert systems in laboratory medicine: A practical consultative application. *Journal of Medical Systems, 13*(2), 103-109.

Punjabi, N., Shade, D., & Wise, R. A. (1998). Correction of single-breath helium lung volumes in patients with airflow obstruction. *Chest, 114*(3), 907.

Quanjer, P. (2009). The GOLD Controversy. Retrieved 20th January, 2011, from http://www.spirxpert.com/controversies/controversy.html

Quanjer, P., Tammeling, G., Cotes, J., Pedersen, O., Peslin, R., & Yernault, J. (1993). Lung volumes and forced ventilatory flows. *The European Respiratory Journal. Supplement, 6*(16), 5-40.

Quinlan, J. R. (1979). Discovering rules by induction from large collections of examples *Expert Systems in the Micro Electronic Age*. Edinburgh: Edinburgh University Press.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81-106.

Quinlan, J. R. (1987). Simplifying Decision Trees. *International Journal of Man-Machine Studies, 27*(3), 221-234.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Sydney: Morgan Kaufmann Publishers Inc.

Ramadan, Z., Compton, P., Preston, P., Le-Gia, T., Chellen, V., Mulholland, M., et al. (1998). *From Multiple Classification RDR to Configuration RDR.* Paper presented at the 11th Knowledge Acquisition for Knowledge-Based Systems Workshop, Banff, Canada.

Richards, D. (1998). *Ripple down rules with formal concept analysis: A comparison to personal construct psychology.* Paper presented at the 11th Knowledge Acquisition for Knowledge-Based Systems Workshop, Banff, Canada.

Richards, D. (2001). Combining cases and rules to provide contextualised knowledge based systems. *CONTEXT 2001, Lecture Notes in Artifical Intelligence, 2116*, 465-469.

Richards, D. (2009). A social software/Web 2.0 approach to collaborative knowledge engineering. *Information Sciences, 179*(15), 2515-2523.

Richards, D., & Busch, P. (2003). Acquiring and Applying Contextualised Tacit Knowledge. *Journal of Information and Knowledge Management, 2*, 179-190.

Richards, D., & Compton, P. (1997a, June 18-20). *Combining Formal Concept Analysis and Ripple Down Rules to Support the Reuse of Knowledge.* Paper presented at the Proceedings Software Engineering Knowledge Engineering (SEKE) 97, Madrid.

Richards, D., & Compton, P. (1997b). *Finding Conceptual Models to Assist Validation*.

Richards, D., & Compton, P. (1997c). Uncovering the conceptual models in ripple down rules. *Conceptual structures: Fulfilling Peirce's dream*, 198-212.

Richards, D., & Compton, P. (1999). *Revisiting Sisyphus I — An Incremental Approach to Resource Allocation using Ripple Down Rules.* Paper presented at the 12th Knowledge Acquisition for Knowledge-Based Systems Workshop, Banff, Canada.

Richards, D., & Vazey, M. (2005). *Closing the Gap Between Different Knowledge Sources and Types in the Call Centre.* Paper presented at the Australasian Conferences on Information Systems (ACIS) 2005, Sydney, Australia.

Rissland, E., & Skalak, D. (1989). *Combining Case-Based and Rule-Based Reasoning: a heuristic approach.* Paper presented at the 11th International Joint Conference on Artificial Intelligence, Detroit.

Roberto J. Bayardo, J., & Agrawal, R. (1999). Mining the most interesting rules *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 145-154). San Diego, California, United States: ACM Press.

Roca, J., Rodriguez-Roisin, R., Cobo, E., Burgos, F., Perez, J., & Clausen, J. (1990). Single-breath carbon monoxide diffusing capacity prediction equations from a Mediterranean population. *American Review of Respiratory Diseases, 141*(4 Pt 1), 1026.

Roddick, J., Fule, P., & Graco, W. (2003). Exploratory medical knowledge discovery: Experiences and issues. *ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) Explorations, 5*(1), 94-99.

Rova, M., Haataja, R., Marttila, R., Ollikainen, V., Tammela, O., & Hallman, M. (2004). Data mining and multiparameter analysis of lung surfactant protein genes in bronchopulmonary dysplasia. *Human molecular genetics, 13*(11), 1095.

Ruppel, G. L. (1994). *Manual of Pulmonary Function Testing* (7 ed.). St Louis: Mosby.

Schank, R. (1980). Language and memory. *Cognitive Science, 4*(3), 243-284.

Sester, M. (2000). Knowledge acquisition for the automatic interpretation of spatial data. *International Journal of Geographical Information Science, 14*(1), 1-24.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379/423.

Shaw, M. L. G., & Gaines, B. R. (1989). Comparing conceptual structures: consensus, conflict, correspondence and contrast. *Knowledge acquisition, 1*(4), 341-363.

Shaw, M. L. G., & Woodward, J. B. (1988). Validation in a knowledge support system: construing and consistency with multiple experts. *International Journal of Man-Machine Studies, 29*(3), 329-350.

Shortliffe, E. (1974). *A rule-based computer program for advising physicians regarding antimicrobial therapy selection.* PhD Thesis, Stanford University, Stanford, California.

Silberschatz, A., & Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering, 8*(6), 970-974.

Simon, H., & Lea, G. (1974). Problem solving and rule induction: A unified view. *Knowledge and cognition*, 105-127.

Singh, P. K. (2006). *Knowledge-based Annotation of Medical Images.* PhD Thesis, University of New South Wales, Sydney.

Sinha, A. P., & Zhao, H. (2008). Incorporating domain knowledge into data mining classifiers: An application in indirect lending. *Decision Support Systems, 46*(1), 287-299.

Smith, H., Irvin, C., & Cherniack, R. (1992). The utility of spirometry in the diagnosis of reversible airways obstruction. *Chest, 101*(6), 1577.

Snow, M., Fallat, R., Tyler, W., & Hsu, S. (1988). Pulmonary Consult: Concept to application of an expert system. *Journal of Clinical Engineering, 13*(3).

Stansfield, S. (2009). ANGY: A rule-based expert system for automatic segmentation of coronary vessels from digital subtracted angiograms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*(2), 188-199.

Steeves, L. (1965). The Need for Continuing Medical Education. *Canadian Medical Association Journal, 92*(14), 758.

Steinberg, T., Wang, Y., Ford, J., & Makedon, F. (2008). A Medical Data Collection System for Sharing Data with Outside Collaborators. *Relation, 10*(1.104), 8875.

Stoyanova, N., & Kommers, P. (2002). Concept Mapping as a Medium of Shared Cognition in Computer-Supported Collaborative Problem Solving. *Journal of Interactive Learning Research*, 111-134.

Stritt, M., & Garland, J. (2009). Effects of Obesity on Lung Volumes and Spirometry. *American Journal of Respiratory and Critical Care Medicine, 179*(1 MeetingAbstracts), A5526.

Stumme, G., Wille, R., & Wille, U. (1998). Conceptual knowledge discovery in databases using formal concept analysis methods. *Principles of Data Mining and Knowledge Discovery*, 450-458.

Subbarao, P., Lebecque, P., Corey, M., & Coates, A. (2004). Comparison of spirometric reference values. *Pediatric pulmonology, 37*(6), 515-522.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Paper presented at the National Academy of Sciences of the United States of America 2005.

Suryanto, H., & Compton, P. (2000). *Learning Classification taxonomies from a classification knowledge based system.* Paper presented at the Proceedings of Workshop on Ontology Learning at ECAI-2000, Berlin.

Suryanto, H., & Compton, P. (2001). *Discovery of ontologies from knowledge bases.* Paper presented at the 1st International Conference on Knowledge Capture, British Colombia, Canada.

Suryanto, H., Richards, D., & Compton, P. (2002). *The automatic compression of multiple classification ripple down rule knowledge based systems: preliminary experiments.* Paper presented at the 3rd International Conference on Knowledge-Based Intelligent Information Engineering Systems, Adelaide, Australia.

Szarfman, A., Machado, S., & ONeill, R. (2002). Use of Screening Algorithms and Computer Systems to Efficiently Signal Higher-Than-Expected Combinations of Drugs and Events in the US FDAs Spontaneous Reports Database. *Drug Safety, 25*(6), 381-392.

Tan, P. N., & Kumar, V. (2001). *Interestingness Measures for Association Patterns: A Perspective*: Army High Performance Computing Research Center, University of Minnesota.

Tapscott, D. (1998). *Growing up digital: The Rise of the Net Generation*. New York: McGraw-Hill.

Thearling, K. (1998). *Some thoughts on the current state of data mining software applications.* Paper presented at the Keys to the Commercial Success of Data Mining: Knowledge Discovery in Databases (KDD) 98, New York.

Thomson, R. (2009). PUFF: Expert System for the Interpretation of Pulmonary Function Tests for Patients with Lung Disease. Retrieved 20th January, 2010, from http://www.openclinical.org/aisp_puff.html

Towell, G. G., & Shavlik, J. (1994). Knowledge-based artificial neural networks. *Artificial Intelligence, 70*(1), 119-165.

Towell, G. G., & Shavlik, J. W. (1993). Extracting refined rules from knowledge-based neural networks. *Machine Learning, 13*(1), 71-101.

Tsumoto, S. (1998). *Modelling medical diagnostic rules based on rough sets.* Paper presented at the First International Conference on Rough Sets and Current Trends in Computing, Warsaw, Poland.

Tsumoto, S. (2004). Mining diagnostic rules from clinical databases using rough sets and medical diagnostic model. *Information Sciences, 162*(2), 65-80.

Tsumoto, S., & Tanaka, H. (1996). *Automated Discovery of Medical Expert System Rules from Clinical Databases based on Rough Sets.* Paper presented at the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon.

Tukey, J. W. (1977). *Exploratory Data Analysis*. London: Addison-Wesley.

van der Lee, I., van Es, H. W., Noordmans, H. J., van den Bosch, J. M. M., & Zanen, P. (2006). Alveolar Volume Determined by Single-Breath Helium Dilution Correlates with the High-Resolution Computed Tomography-Derived Nonemphysematous Lung Volume. *Respiration, 73*(4), 468-473.

van Ganse, W., Comhaire, F., & van der Straeten, M. (1970). Residual volume determined by single breath dilution of helium at various apnoea times. *Scandinavian Journal of Respiratory Diseases, 51*(2), 73.

Vazey, M. (2006). Stochastic foundations for the case-driven acquisition of classification rules. *Managing Knowledge in a World of Networks*, 43-50.

Vazey, M., & Richards, D. (2005). *Troubleshooting at the Call Centre: A Knowledge-based Approach.* Paper presented at the Artificial Intelligence and Applications 2005, Innsbruck, Austria.

Vazey, M., & Richards, D. (2006). Evaluation of the FastFIX prototype 5Cs CARD system. *Advances in Knowledge Acquisition and Management*, 108-119.

Walser, R., & McCormick, B. (1977). *A System for Priming a Clinical Knowledge Base.* Paper presented at the American Federation of Information Processing Societies Conference '77.

Wanger, J., Clausen, J., Coates, A., Pedersen, O., Brusasco, V., Burgos, F., et al. (2005). Standardisation of the measurement of lung volumes. *European Respiratory Journal, 26*(3), 511.

Waterman, D. (1970). Generalization learning techniques for automating the learning of heuristics* 1. *Artificial Intelligence, 1*(1-2), 121-170.

Watson, I., & Marir, F. (1994). Case-Based Reasoning: A Review. *The Knowledge Engineering Review, 9*(4), 327-354.

Wille, R. (1982). *Restructuring Lattice Theory: an approach based on hierarchies of concepts.* Paper presented at the Ordered Sets, Dordrecht, Reidel.

Wille, R. (1989). *Knowledge acquisition by methods of formal concept analysis.* Paper presented at the Data Analysis, Learning Symbolic and Numeric Knowledge, New York.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques* (2 ed.). San Francisco: Morgan Kaufmann.

Yamaguti, T., & Kurematsu, M. (1993). *Legal knowledge acquisition using case-based reasoning and model inference.* Paper presented at the 4th International Conference on Artificial Intelligence and Law, Amsterdam, The Netherlands.

Zadeh, L. A. (1979). *Approximate reasoning based on fuzzy logic.* Paper presented at the 6th lnternational Conference on Artificial Intelligence, Tokyo, Japan.

Zagzebski, L. (1999). What is Knowledge? In J. Greco & E. Sosa (Eds.), *Epistemology* (pp. 92–116). Oxford: Blackwell Publishers.

Zhang, C., Yu, P. S., & Bell, D. (2009). Domain-driven data mining. *IEEE Transactions on Knowledge and Data Engineering, 21*(2), 301.

Zhang, D., Zhou, L., & Nunamaker Jr, J. F. (2002). A knowledge management framework for the support of decision making in humanitarian assistance/disaster relief. *Knowledge and Information Systems, 4*(3), 370-385.

Appendix A – Additional Data Analysis Tables and Figures

Attribute	Expected Mean	Actual Mean	Std Deviations
FEV_1 % Δ	5.59%	15.71%	1.3
FEV_1 Δ	0.12	0.26	0.8
FEV_1 % of predicted pre-BD	79.31%	57.46%	0.9
FEV_1 % of predicted post-BD	84.24%	65.69%	0.7
FEV_1/FVC pre-BD	0.71	0.59	0.9
FEV_1/FVC post-BD	0.74	0.57	1.2
PEF % of predicted pre-BD	93.08%	66.89%	1.2
PEF % of predicted post-BD	95.37%	75.44%	0.9
FEF_{25} pre-BD	4.62	2.29	0.9
FEF_{25} post-BD	5.01	2.66	0.8
$FEF_{25-75\%}$ % of pred. pre-BD	77.2%	44.63%	0.9
$FEF_{25-75\%}$ % of pred. post-BD	86.56%	46.87%	1.1
FEF_{50} pre-BD	3.4	1.9	0.9
FEF_{50} post-BD	3.76	2.08	1
RV % of predicted	108.69%	136.86%	0.7
FRC % of predicted	101.65%	118.4%	0.5
V_A/TLC	0.86	0.78	0.5
D_LCO uncorrected % of pred.	84.13%	70.48%	0.6
TLC – V_A	0.87	1.44	0.6

Table A-1: Attributes indicated as related to the *FVC Reversibility* class

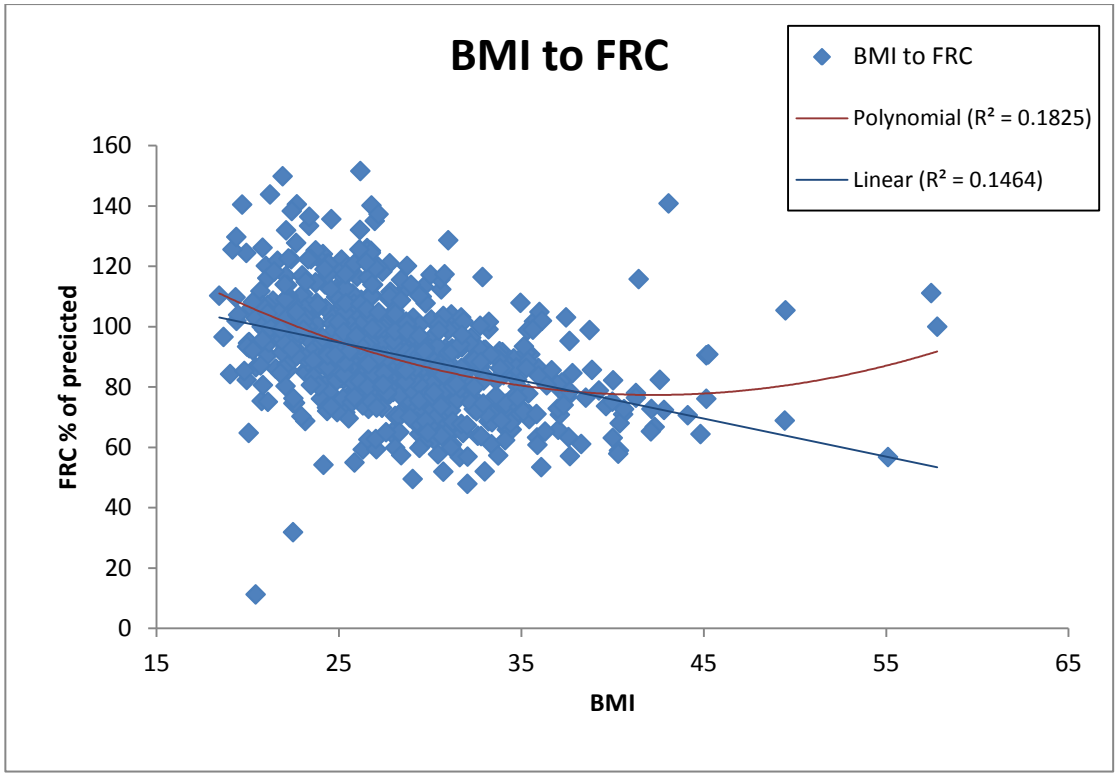


Figure A-1: BMI to FRC comparison, for all weight groups

Appendix B – Pre and Post Knowledge Acquisition Questionnaires

Pre Questionnaire:

1. How would you rate your level of experience with interpreting lung function reports?

5 (Highly experienced) **4** (Experienced) **3** (Some) **2** (Little) **1** (Very little experience)

At a rough estimate, how many lung function reports have you worked with?

_____

2. How would you rate your confidence in interpreting lung function reports?

5 (Very confident) **4** (Confident) **3** (Don't know) **2** (Not confident) **1** (Very unconfident)

Comments:

_____
_____
_____
_____
_____

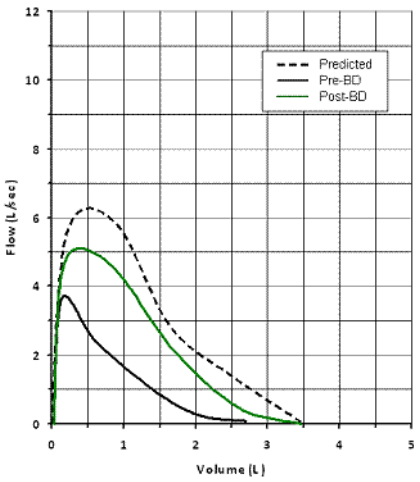
3. What conclusions would you draw about the attached report, and why? Be as detailed in your answer as you can.

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

Patient ID:	482	Age:	35	Sex:	F	Date:	2005	Weight:	104	
Test ID:	482	Smoker:	N	Pack Years:		Ethnicity:	Cauc	Height:	164	BMI: 38

	Pre - Bronchodilator			Post - Bronchodilator		
	Actual	Predicted	% Predicted	Actual	% Predicted	% Change
Spirometry						
FEV1 (L)	1.57	2.96	53.0	2.34	79.1	49.0
FVC (L)	2.68	3.49	76.8	3.46	99.1	29.1
FEV1/FVC	0.59	0.85	69.4	0.68	80.0	
PEF (L/sec)	3.57	6.43	55.5	5.1	79.3	
FEF 25% (L/sec)	2.46	6.05	40.7	4.89	80.8	40.2
FEF 75% (L/sec)	0.25	1.79	14.0	0.46	25.7	11.7
FEF 25-75% (L/sec)	0.69	3.35	20.6	1.35	40.3	19.7
Lung Volumes						
SVC (L)	2.75	3.49	78.8			
IC (L)	2.62	2.17	120.7			
ERV L)	0.13	1.32	9.9			
RV (L)	2.49	1.51	164.9			
TLC (L)	5.24	5	104.8			
RV/TLC (%)	0.48	0.29	165.5			
FRC (L)	2.62	2.83	92.6			
Diffusion						
DLCOcor (ml/min/mmHg)	28.55	26.5	107.7			
VA (L)	4.25	4.85	87.6			
TLC - VA (L)	0.99					
DLCOunc/VA (ml/min/mmHg/L)	6.79	5.46	124.4			
DLCOcor/VA (ml/min/mmHg/L)	6.72	5.46	123.1			
Blood Gases						
Hgb	15					

Flow Volume Loop



Post Questionnaire:

1. How easy to use was the software interface?

5 (Very easy) **4** (Fairly easy) **3** (Don't know) **2** (Difficult) **1** (Very difficult)

2. If it occurred, how useful were the indications of conflicting rules?

5 (Very helpful) **4** (Helpful) **3** (Don't know) **2** (Unhelpful) **1** (Completely unhelpful)

3. For the statistical information (shown on the right of the screen when making a rule):

a) Did you understand what information was being expressed?

Yes **No**

b) How helpful was the information when making a rule?

5 (Very helpful) **4** (Helpful) **3** (Don't know) **2** (Unhelpful) **1** (Completely unhelpful)

c) How often did the information influence your rule making decisions?

5 (Always) **4** (Mostly) **3** (Don't know) **2** (Sometimes) **1** (Never)

d) Were there any further information or statistics you would like to have seen? Do you have any further comments?

_____

_____

_____

_____

4. How would you now rate your confidence in interpreting lung function reports?

5 (Very confident) **4** (Confident) **3** (Don't know) **2** (Not confident) **1** (Very unconfident)

5. Do you feel that you learned more about lung function through this process?

Yes **No**

If so, what do you feel that you have learned?

_____

_____

_____

_____

_____

_____

Continued overpage…

6. Would you use this (or a similar) system again?

Yes **No**

Comments:

Appendix C – Ethics Consent Form and Participant Information Sheet

PARTICIPANT INFORMATION SHEET (PROFORMA)
SOCIAL SCIENCE/ HUMANITITES
RESEARCH

Knowledge Consolidation and Development in the Health Domain

Invitation
You are invited to participate in a research study into developing an online system which can assist in performing medical research, and can compare and combine medical knowledge.

The study is being conducted by Tristan Ling, a PhD candidate within the School of Computing and Information Systems at the University of Tasmania, with supervisors Associate Professor Byeong Ho Kang from the School of Computing and Information Systems, Associate Professor Justin Walls from the School of Medicine, and Associate Professor David P Johns from the Menzies Research Institute and School of Medicine.

1. 'What is the purpose of this study?'
In broad terms, the research aims of this PhD candidature are to:

- Analyse and compare how human experts interpret patient cases
- Develop an online computerised system for providing 'intelligent' interpretation of patient case data
- Provide an online interface for human experts to examine archived data and test hypotheses against the stored lung function knowledge and data

.

2. 'Why have I been invited to participate in this study?'
You are eligible to participate in this study because of your expertise in the field of lung function test interpretation.

4. 'What does this study involve?'
In order to reach any of these goals, the online system needs to learn how experts decide on their classifications. The only way it can learn this is for human experts to interact with it and "teach" it. In training the system all that is required is to perform the normal tasks involved in interpreting cases – and when the system disagrees with the experts' conclusions, to have the expert create a rule that defines how they reached those conclusions. The rules are very simple to create: just select which values were used to make the decision. For example, you may define that the case shows obstruction

because the FEV1 to FVC ratio is less than 80% of the predicted value. The incidences of the knowledge base being incorrect will get exponentially less the more rules that are entered.

It is important that you understand that your involvement is this study is voluntary. While we would be pleased to have you participate, we respect your right to decline. There will be no consequences to you if you decide not to participate, and this will not affect your treatment / service. If you decide to discontinue participation at any time, you may do so without providing an explanation. All information will be treated in a confidential manner, and your name will not be used in any publication arising out of the research. All of the research will be kept on a physically and electronically secured web server within the School of Computing and Information Systems.

5. Are there any possible benefits from participation in this study?

If the study is successful the system will provide a tool for assisting medical researches in testing hypotheses against a large body of patient data and also against the compiled knowledge of experts from across the field. All participants will be invited to make use of the tool in their own research. The system will also be a capable and reliable expert at interpreting lung function test results, with a wide range of expertise.

6. Are there any possible risks from participation in this study?

There are no specific risks anticipated with participation in this study.

7. What if I have questions about this research?

If you would like to discuss any aspect of this study please feel free to contact Tristan Ling on ph 03 6226 2910 or at Tristan.Ling@utas.edu.au.

This study has been approved by the Tasmanian Social Science Human Research Ethics Committee. If you have concerns or complaints about the conduct of this study should contact the Executive Officer of the HREC (Tasmania) Network on (03) 6226 7479 or email human.ethics@utas.edu.au. The Executive Officer is the person nominated to receive complaints from research participants. You will need to quote [H10834].

Thank you for taking the time to consider this study.
If you wish to take part in it, please sign the attached consent form.
This information sheet is for you to keep.

CONSENT FORM

Title of Project: **Knowledge Consolidation and Development in the Health Domain**

1. I have read and understood the 'Information Sheet' for this project.
2. The nature and possible effects of the study have been explained to me.
3. I understand that the study involves accessing a website and interpreting lung function test results, and that my decisions in doing this will be recorded for future analysis and use.
4. I understand that all research data will be securely stored on the University of Tasmania premises for five years [or at least five years], and will then be destroyed [or will be destroyed when no longer required].
5. Any questions that I have asked have been answered to my satisfaction.
6. I agree that research data gathered from me for the study may be published provided that I cannot be identified as a participant.
7. I understand that the researchers will maintain my identity confidential and that any information I supply to the researcher(s) will be used only for the purposes of the research.
8. I agree to participate in this investigation and understand that I may withdraw at any time without any effect, and if I so wish, may request that any data I have supplied to date be withdrawn from the research.

Name of Participant: _____

Signature: _____ Date: _____

Statement by Investigator

☐ I have explained the project & the implications of participation in it to this volunteer and I believe that the consent is informed and that he/she understands the implications of participation

If the Investigator has not had an opportunity to talk to participants prior to them participating, the following must be ticked.

☐ The participant has received the Information Sheet where my details have been provided so participants have the opportunity to contact me prior to consenting to participate in this project.

Name of Investigator Tristan Ling

Signature of Investigator

Name of investigator _____

Signature of investigator _____ Date _____