5

Building Clinical Trust in Automated Knowledge Acquisition

Anna Shillabeer
Carnegie Mellon University
Australia

1. Overview

The aim of this chapter is to describe the process of medical knowledge acquisition from a historical context and to define the requirements for producing knowledge which is able to be trusted and applied in a clinical setting. This is related to modern data mining approaches which do not yet adequately address these requirements. This is believed to be the most critical issue in the acceptance of data mining in the medical domain. The chapter will discuss how data mining *can* address these needs and will provide discussion of a technical solution to the stated issues. Overall this chapter aims to demonstrate that the individual needs of all medical professionals can be addressed and that data mining can be a valuable tool in the diagnostic and decision making toolkit. It will also empower medical professionals to take a greater role in the development of such systems by providing a checklist of features to include and pitfalls to avoid, thus ensuring greater success in future systems.

2. Introduction

2.1 Clinical data mining context

While acceptance of data mining technologies is growing, progress has been slow and it is not yet an integrated part of the medical data analysis toolkit. Many reasons have been documented for this but the primary issues are two fold; the decision making and knowledge acquisition processes of the medical domain are not adequately reflected in the technologies available and the systems are too often built to suit the specific analytical needs of an individual user. Whilst this has enabled the application of the technology in specific scenarios, it has resulted in the development of tools which cannot be utilised outside of the specific purpose for which they were built. These issues serve to limit the exposure, applicability and trust of data mining systems to the medical domain.

Data mining researchers have long been concerned with the application of tools to facilitate and improve data analysis on large, complex data sets for the purpose of knowledge acquisition. The current challenge is to make data mining and knowledge discovery systems applicable to a wider range of domains, among them medicine. Early work was performed over transactional, retail based data sets, but the attraction of finding previously unknown knowledge from the increasing volume of data collected from the medical domain is an emerging area of interest and specialisation. This chapter is primarily concerned with

Source: Data Mining in Medical and Biological Research, Book edited by: Eugenia G. Giannopoulou,
ISBN 978-953-7619-30-5, pp. 320, December 2008, I-Tech, Vienna, Austria

defining the manner in which new knowledge is acquired, what constitutes acceptability in new knowledge for the medical domain and how this can be measured and automated through the application of data mining technologies. There is a growing body of work which aims to qualify and define a process for the discovery of new knowledge accross a range of domains, however this work has not focused on the unique needs of the medical domain and hence the domain has remained relatively untouched by the advances in data mining and knowledge discovery technology.

The primary challenge presented by medicine is to develop a technology that can apply a trusted knowledge acquisition process to reveal data patterns in the form of hypotheses which are based on measures that can be relied upon in medical health research and tested in a clinical environment (Ashby & Smith 2002). Due to the broad nature of medical informatics and the diversity of professional roles in medicine, the requirement is to find a solution that is both flexible enough to address the unique and varied data management and analysis requirements within the domain, whilst being specific enough to address the individual needs of an equally broad range of users. To date, automated data analysis systems have been developed for a particular role or field of investigation using a specific and relatively homogenous data set and are not transferable to other professional roles, fields or data sets within the domain. The primary requirements are therefore the provision of a methodology and system to facilitate the acquisition of knowledge which conforms to the requirements of the domain and, the production of statistically valid hypotheses which are appropriately targeted to the role of an individual user and which can provide a sound foundation for further research or clinical trials.

Data mining in medicine is most often used to complement and expand the work of the clinician and researcher by qualifying or expanding knowledge rather than providing new knowledge as is the trend in other domains. Very little health data mining is purely exploratory and hence the technology is generally not applied to provide novel knowledge i.e. to identify previously unknown patterns hidden within the data. One of the difficulties in providing new knowledge in the health domain is the need to sufficiently cross reference and validate the results. It is not sufficient to provide a standard rule in the form of A gives B in the presence of C without substantiating the information held therein. This information could already be known, may be contrary to known medical facts, incomplete due to missing attributes, may not be statistically valid by trusted measures or may simply not relate to the specialisation of the user and is therefore contextually irrelevant.

The barriers to the application of data mining to medical data can be generalised as follows:

1. The low level of flexibility in data mining systems and the need for medical analytical processes to adapt to data mining methodologies rather than data mining adapting to the needs of medicine,
2. The lack of opportunity for incorporating subjectivity when mining medical data,
3. The production of patterns in a technical language and format that are often not understandable or applicable in a clinical setting,
4. The broad range of users and analytical variance in medicine,
5. The production of too many irrelevant results, requiring a high level of user interpretation to discriminate those that are truly useful.

This chapter presents a flexible solution to these issues through discussion of a novel process to more closely reflect clinical medical processes in the technical data mining process. This is achieved through the development of two complementary systems; a

hypothesis engine and an Automated DAta Pattern Translator (ADAPT) that provide a mechanism for the translation of technical data mining outputs into a language which can be better understood and accepted by medical professionals. As an integrated unit, the hypothesis engine and ADAPT are able to facilitate greater access to mining technologies, and the ability to apply some of the more complex mining technologies by all medical users without the risk of producing irrelevant or incomprehensible outputs.

The health domain is a myriad of complexity and standardised data mining techniques are often not applicable (Cios, 2002, Imberman & Domanski 2002; Hagland, 2004) hence the need for a deeper analysis of the potential for data mining in the medical domain which in turn requires knowledge of the process of medical knowledge acquisition and how the data mining technologies can facilitate this process. The remainder of this chapter aims to reduce this knowledge gap through a discussion of the processes of knowledge acquisition and diagnostic decision making in the medical domain and of the potential for novel data pattern evaluation methods to augment and automate these processes. Discussion of a collaborative two part solution developed through a merging of theories from data mining, medicine and information retrieval theory will be presented, together with experimental results to demonstrate the application of these solutions to the issues identified in this section.

2.2 The history of medical knowledge acquisition

History has taught us that standardised scientific processes have been followed for centuries to ensure that only trusted, proven knowledge is applied in a clinical setting. The process for defining what and how new medical knowledge is trusted does not readily correlate with current data mining processes. However where the two are combined the rate of acceptance of outputs is higher than in those where data mining process alone is enforced upon the medical knowledge acquisition process. To understand the similarities and differences between the two processes both processes must first be defined.

Throughout recorded history there has been debate over what constitutes knowledge and therefore what constitutes proof of knowledge. Early practitioners of medical science, such as Hippocrates, based their knowledge development in philosophy and their ability "to see with the eye of the mind what was hidden from their eyes" (Hanson, 2006) . By the first century A.D. physicians, such as Galen, were beginning to question the validity and contradictions of Hippocrates work which had stood mostly unopposed since the 5th Century B.C. It is not clear if there was any agreement or understanding of the methods applied by physicians to develop their knowledge base at this time as there was no empirical proof or scientific process documented. Galen was one of the first to suggest that there should be a process for the provision of substantiated evidence to convince others of the value of long held medical beliefs and hence raised the notion of a practical clinical method of knowledge acquisition which combined the Hippocratic concept of hypothesis development through considered thought and a priori knowledge, with clinical observation to evaluate and hence provide proof or otherwise of the hypothesis. This general methodology has survived to the present day and is reflected not only in the provision and acceptance of new knowledge but also in the process of clinical diagnosis.

The historical debate on knowledge acquisition methodologies has primarily focused on three philosophical groups; Methodists, Empiricists and Rationalists. Whilst these three groups are most frequently discussed in a Graeco-Roman context, they were either being

applied or paralleled in various other cultural contexts including India and Islam. All three of these cultural contexts are discussed here briefly to demonstrate the extent and foundations of medical knowledge acquisition debate in the ancient world.

2.2.1 The Graeco-Roman context

- Methodists

The first prominent physician practicing according to the Methodist philosophy was Hippocrates of Cos (460-380 B.C.) who is still referred to as the "Father of Medicine" (Hanson, 2006). It is believed by many that he initiated the production of over 60 medical treatises known as the Hippocratic Corpus. The corpus was written over a period of 200 years and hence had more than one author which is reflected in the sometimes contradictory material contained therein. The body of work was however consistent in its reliance on defining a natural basis for the treatment of illnesses without the incorporation or attribution of magic or other spiritual or supernatural means as had occurred previously. Methodists were defined as those whom attributed disease etiology and treatment primarily to an imbalance in bodily discharges. Illnesses were categorised by whether they represented a withholding of fluids, for example a blister holds water, or an excessive releasing of fluids, for example a weeping eye. This group founded its knowledge on an understanding of the nature of bodily fluids and developed methods for the restoration of fluid levels. They were not concerned with the cause of the imbalance or the effect on the body of the imbalance, only in recognising whether it was an excess or lack of fluid and the method for treating that observation.

- Rationalists

Rationalists believed that to understand the workings of the human body it was necessary to understand the mechanism of illness in terms of where and how it affected the body's functioning (Brieger, 1978). They were not interested in the treatment or diagnosis of illness but focused on understanding and recording the functioning of the living system. Two works are of prominence in this group (Corsans, 1997); the Timaeus by Plato which systematically described the anatomical organisation of the human body and; Historia animalium by Aristotle which discussed further both human and animal anatomy and the links between such entities as the heart and blood circulation. This method of knowledge acquisition was criticised as it effectively removed medicine from the grasp of the average man and moved it into a more knowledge based field where philosophical debate or an observational experiential approach was not deemed sufficient (Brieger, 1978). Essentially Rationalists did not believe in a theory unless it was accompanied by reason. They espoused the requirement for knowledge to be founded on understanding both cause and effect of physical change in the body (Horton, 2000).

- Empiricists

The Empiricists believed that it was not enough to understand how the body works and reacts to illness. They pursued a philosophy which stated that it was necessary to demonstrate the efficacy of treatments and provide proof that a treatment is directly responsible for the recovery of a patient rather than providing academic argument regarding why it should result in recovery. Galen is considered to be one of the earliest empiricists (Brieger, 1978). He was both a medical practitioner and a prolific scholarly writer and is certainly one of the best known and more frequently quoted empiricists. He was particularly interested in testing the theories proposed in the Hippocratic Corpus, especially

given its frequent contradictions. His work was also produced at a time when medicine as a science was evolving from its previous status as a branch of philosophy. In his work Galen argues that "medicine, understood correctly, can have the same epistemological certainty, linguistic clarity, and intellectual status that philosophy enjoyed" (Pearcy, 1985). Empiricists were the first to concentrate on the acquisition of knowledge through demonstrated clinical proof developed through scientific methodologies which provided conclusive statements of cause and effect.

2.2.2 The islamic context

- The Empiricists (Ashab al-Tajarib).

Dr. Mahdi Muhaqqiq was an early 20th century Iranian scholar who wrote texts on many subjects including medical knowledge acquisition throughout the history of Iran. He recorded that the early Empiricists believed that medical knowledge was derived from experience obtained through the use of the senses and that the knowledge is comprised of four types; "incident (ittifaq), intention (iradah), comparison (tashbih) and the adoption of a treatment that was used in another similar case (naql min shay' iki shabihihi)" (Muhaqqiq, 2007).

- Incident - this can either describe a natural event such as a sweat or headache, or an accidental event such as a cut or a broken limb.
- Intention - denotes an event experienced by choice for example taking a cool bath to reduce a fever.
- Comparison - A technique employed by a practitioner whereby he notes that one of the above techniques results in a useful effect which can be applied to other similar presentations. For example applying cold water to reduce localised burning of the skin following the observation that a cool bath can reduce generalised fever or body heat.
- Naql - A technique whereby the physician applies a treatment for a similar presentation in the instance of a presentation which has not been encountered before. An example might be the prescribing of a medication for a previously unencountered infected tooth where that medication had only previously been used for an infection elsewhere in the body.

The empiricists treated a patient through knowledge of that patient's demographics and therefore all patients of a certain age and sex with some similar complaint were treated the same whereas patients of the opposite sex may have been treated differently even though the condition was the same. Their knowledge was based on patient characteristics rather than a specific condition or set of symptoms. Whilst this seems to differ from the Graeco-Roman definition of empiricism, both groups believed that knowledge acquisition occurred through observing or testing the effect of a treatment and producing rules based on what is considered reliable empirical proof rather than conjecture and debate.

- The Dogmatists (Ashab al-Qiyas).

The Dogmatists believed that while scientific belief and knowledge should be derived from experience and observation this should be tempered by the use of thought and considered evaluation (Mohaghegh, 1988). They believed that changes in the bodily functions must be precipitated by some event and that it is necessary to not only understand what these changes are but also what the specific causes of those changes are in order to correctly diagnose and treat any condition. Changes are defined as being of two types (Muhaqqiq, 2007):

- Necessary change - drink reducing thirst. This is a change which is required for normal bodily functioning.
- Unnecessary change - dog bite causing bleeding. This change is not a requirement to aid or enhance bodily wellbeing.

Dogmatists based their treatments upon the nature of the condition rather than the type of patient as seen with the empiricists. The treatments were therefore selected through knowledge of the causes of illness and the effects of those treatments upon the illness or symptoms. This required an understanding of the physical body and the changes that result from illness in a similar manner to that of the Graeco-Roman Rationalists.

- The Methodists (Ashab al-Hiyal).

This group believed in a generalist view of illness and treatment and categorised conditions in terms of the extent to which bodily fluids and wastes are either retained and/or expelled. Treatments were generally natural remedies based upon adjusting the balance between such aspects of life as food and drink, rest and activity, etc. Methodists were not interested in the type of patient or cause and effect of illness and were hence considered to be more prone to error (Muhaqqiq, 2007). This is in direct parallel to the Methodist philosophy discussed in Section 2.2.1.

It has been suggested that in general, Islamic physicians relied primarily upon analogy which reflects their focus on logic in other scholarly areas (Mohaghegh 1988). This has resulted in widespread support for the Dogmatist methods of knowledge acquisition through research and understanding of cause and effect in the human system. However there is still debate between scholars with some believing that Dogmatism alone is the only method of ensuring progress in medical diagnosis and treatment as it is the only method which tries to seek new understanding rather than relying upon past experience or a closed assumption that there is a single cause for all illness (Mohaghegh, 1988). Others prefer to adhere to the Graeco-Roman perspective (developed by Plato) that a combination of experience and analogy is required if a holistic, 'correct' practice of medicine is to be achieved (Muhaqqiq, 2007).

2.2.3 The Indian context

India is not well known for its scientific contributions or texts, however it has a long history in the development of medical knowledge. In the 11th century a Spanish scholar, Said Al-Andalusi, stated that he believed that the Indian people were "the most learned in the science of medicine and thoroughly informed about the properties of drugs, the nature of composite elements and the peculiarities of the existing things" (al-Andalusí, 1991). The reasons for this apparent invisibility of Indian scientific progress may be due to religious debate in India which has frequently negated the influence of scientific explanation instead preferring to rely upon mystical or spiritual beliefs. There are however documented scientific approaches to the development of a body of knowledge regarding medicine from centuries before the texts of Hippocrates and which, although often earlier, discuss similar theories to those presented in the Graeco-Roman texts.

- The Rationalist schools

One of the earliest groups to produce texts concerning the acquisition of knowledge regarding the human state were the Upanishads which were believed to have been written between 1500 and 600B.C. and were concerned with knowledge regarding the spirit, soul and god (Tripod, 2002, Kaul & Thadani 2000). Although these texts were embedded in

mysticism and spirituality they used natural analogy to explain the notion of the soul and god and allowed the expression of scientific and mathematical thought and argument which formed the basis for the emergence of the rationalist period. Early rationalists included the Lokyata, Vaisheshika school and the Nyaya school. These groups espoused a scientific basis for human existence and a non-mystical relationship between the human body and mind. They also developed primitive scientific methodologies to provide "valid knowledge" (Tripod, 2002, Kaul & Thadani 2000).

The Lokyata were widely maligned by Buddhist and Hindu evangelicals as being heretics and unbelievers due to their refusal to ``make artificial distinctions between body and soul'' (Kaul & Thadani 2000). They saw all things in terms of their physical properties and reactions and gave little attention to metaphysical or philosophical argument, preferring to believe only what could be seen and understood. They developed a detailed understanding of chemistry, chemical interactions and relationships between entities. They are also believed to be the first group to document the properties of plants and their uses, this provided an elementary foundation for all pharmaceutical knowledge which followed.

The primary input of the Vaisheshika school toward the progression of human knowledge was their development of a process for classification of entities in the natural world and in their hypothesis that all matter is composed of very small particles with differing characteristics (Tripod, 2002). Their theory stated that particles, when combined, give rise to the wide variety of compounds found upon the earth and allowed them to be classified by the nature of the particles from which they were formed. This school also introduced the notion of cause and effect through monitoring and understanding temporal changes in entities. The importance of this work lay in the application of a methodology for identification and classification of relationships between previously unconnected entities. This early recognition of the need for a documented scientific process provided a mechanism for the schools which followed to present substantiated proof of evidence for theories in the sciences including physics, chemistry and medicine.

The Nyaya school further developed the work of the Vaisheshika school by continuing to document and elaborate a process for acquiring valid scientific knowledge and determining what is true. They documented a methodology consisting of four steps (Tripod, 2002):

- Uddesa was a process of defining a hypothesis.
- Laksan was the determination of required facts "through perception, inference or deduction".
- Pariksa detailed the scientific examination of facts.
- Nirnaya was the final step which involved verification of the facts.

This process would result in a conclusive finding which would either support or refute the original hypothesis.

The Nyaya school also developed definitions for three non scientific pursuits or arguments which were contrary to the determination of scientific truth but which were often applied by others to provide apparent evidence for theories or knowledge (Tripod, 2002, Kaul & Thadani, 2000). These included jalpa to describe an argument which contained exaggerated or rhetorical statements or truths aimed at proving a point rather than seeking evidence for or against a point; vitanda which aimed to lower the credibility of another person and their theories generally through specious arguments; and finally chal, the use of language to confuse or divert the argument.

Further to this again a set of five 'logical fallacies' were developed:

- savyabhichara - denotes the situation where a single conclusion is drawn where there could be several possible conclusions,
- viruddha - where contradictory reasoning was applied to produce proof of the hypothesis,
- kalatita - where the result was not presented in a timely manner and could therefore be invalidated,
- sadhyasama - where proof of a hypothesis was based upon the application of another unproven theory, and
- prakaranasama - where the process simply leads to a restating of the question.

These concepts were unique in their time but many remain applicable in modern scientific research.

This section has demonstrated that the quest for new medical knowledge and a deeper understanding of the human system is not a recent initiative but in fact one which has its foundations up to four centuries B.C. While there were several distinct cultural groups all were primarily concerned with defining the most reliable methodology for evaluating what knowledge could be trusted and applied clinically. The Graeco-Roman and Islamic practitioners were concerned with the means by which evidence was obtained and the Indians were more concerned with methods for proving the validity of knowledge after it had been discovered. Both of these foci remain topics of debate in the 21st century and as late as 1997 a report was published by the International Humanist and Ethical union regarding trusted versus untrusted clinical practices and the requirement for proof of the benefits of medical treatments. The opening of a Mantra Healing Centre at the Maulana Azad Medical College in New Delhi was described as "ridiculing the spirit of inquiry and science" through its application of "sorcery and superstition in their rudest form" (Gopal, 1997). The report did not however argue that there was no worth in mantra healing but that there was no proof of worth as per the requirements of the still flourishing rationalist opinion. The debate on what is trusted and clinically applicable knowledge forms the focus of this chapter as it investigates the application of new knowledge acquisition tools and aims to identify best practice procedures for automating the acquisition of new medical knowledge.

2.3 Non-scientific knowledge acquisition

History has shown that the acquisition of much currently accepted medical knowledge was based on serendipity or chance accompanied by a strong personal belief in an unproven hypothesis. Further to this, much knowledge was acquired through a process which directly contradicts accepted scientific practice. Whilst there was usually a scientific basis to the subsequent development of proof this was often produced through a non traditional or untrusted application of scientific processes. Unfortunately this frequently resulted in lengthy delays in acceptance of the work. The following list provides a range of such breakthroughs over the past 250 years which can be attributed to little more than chance, tenaciousness and the application of often radical methods to obtain proof.

1. James Lind (1716-1794) (Katch, 1997). Based upon an unsubstantiated personal belief that diet played a role in the development of scurvy on naval vessels, Dr Lind performed limited randomised trials to provide proof and then published his Treatise on the Scurvy which is still relevant to this day.
2. Edward Jenner (1749-1823) (Sprang, 2002). During his apprenticeship Dr Jenner overheard a milkmaid suggest that those who have had cowpox can not contract

smallpox. He then tested the theory by infecting a young boy sequentially with each pathogen and as a result created the concept of a vaccine and initiated the global eradication of smallpox.

3. John Snow (1813-1854) (Ucla, 2002, BBC, 2004). Dr Snow believed, without any direct evidence, that the transmission of viral agents was possible through contaminated water. In 1854 he applied the theory and provided an answer to the cholera epidemic.
4. Alexander Fleming (1881-1955) (Page, 2002). Dr Fleming stumbled upon a discarded culture plate containing a mould which was demonstrated to destroy staphylococcus. The mould was isolated and became the active ingredient in penicillin based antibiotics.
5. Henri Laborit (1914-1995) (Pollard, 2006). During his ward visits Dr Laborit noticed that patients given an antihistamine named promethazine to treat shock not only slept but reported pain relief and displayed a calm and relaxed disposition leading to the development of medications to treat mental disorders including schizophrenia.
6. Robert Edwards and Patrick Steptoe (1925 - , 1913-1988) (Swan, 2005, Fauser & Edwards, 2005). These doctors were the first men to deliver a baby through in-vitro fertilisation after 20 failed attempts and great ethical debate following a lack of proof in animal subjects.
7. Barry J Marshall (1951-) (Marshall, 1998). Dr Marshall worked against accepted medical knowledge to provide proof of the bacterial agent, Heliobacter Pylori, as the cause of stomach and duodenal ulcers. So strong was the opposition to initial clinical testing of the theory he resorted to using himself as the test subject.

Whilst each of these examples provided wide reaching benefits to human health and contributed significantly to the body of medical knowledge in some cases, they would not have been possible if only standardised scientific methodologies had been applied using only trusted traditional processes. This demonstrates that there is often a need to do things differently and not only apply what is comfortable and safe to enable the acquisition of knowledge, although there is always a requirement to provide substantiated proof and an argument based upon scientific principles. The applicability of this notion is particularly relevant to this chapter which focuses on the application of new techniques and technologies which have demonstrated an ability to provide an important impetus to the acquisition of knowledge in other domains and which have not been demonstrated to be detrimental to the process in the medical domain. However, the same proof of hypothesis hurdles must be overcome and an equally strong argument and testing methodology must be provided for the resulting knowledge to be accepted. Throughout history the same quality of evidence has been required and the omission of this evidence has often resulted in decades of latency between hypothesis statement and the generation of conclusive evidence in support (or otherwise) of that hypothesis.

Regardless of the methodology for producing the evidence required for knowledge acquisition, the above examples all had to fulfil a number of further requirements prior to the acquired knowledge being accepted. These requirements are summarised following:

1. Replication of results.
2. Non contradictory results.
3. Scientifically justified theories and hypotheses.
4. Ethical methodologies and measures.
5. Results demonstrated to be representative of the population.
6. Results derived from sufficient numbers of cases.
7. Publicly documented processes and results.

2.4 The application of data mining

Medical history has recorded many instances of the manual use of data mining techniques resulting in medical breakthroughs crucial to the preservation of thousands of human lives if not entire populations. Over centuries medical professionals have (often unknowingly) employed the same scientific analytical methods to data as are applied during data mining in order to develop hypotheses or to validate beliefs. Whilst these techniques have been applied in a simplistic form they clearly demonstrate the applicability of the founding principles of data mining to medical inquiry and knowledge acquisition. A number of the examples discussed in Section 2.3 are used to demonstrate this.

- Data sampling - James Lind (Katch, 1997). Dr Lind performed small randomised trials to provide proof of the cause of scurvy. In his position as Naval doctor he was able to test his theories on the crews of the vessels he sailed on, however without documented proof it was not possible to test the entire navy en mass. Developing sufficient proof in this manner was a lengthy process and it was 50 years before the British Admiralty accepted and applied his theories, a delay which cost the lives of many sailors. Lind's process shows the use of examining subsets of the population, being able to clearly identify the variant in the knowledge gained and then substantiating that knowledge by testing on similar populations to ensure the finding is representative is a suitable technique for hypothesis testing and knowledge substantiation.
- Association rules and support and confidence heuristics - Edward Jenner (Sprang, 2002). Following development of a hypothesis from the knowledge that milkmaids were less likely than members of the general population to develop small pox due to their increased contact with cow pox, Jenner conducted further tests over a period of 25 years to validate the relationship and publish his findings. This work demonstrates the use of the concept of support through Jenner's realisation that there was a frequently occurring and previously unknown pattern in a data set or population. That pattern was subsequently tested to provide confidence levels by showing that contracting cowpox almost always results in an inability to contract smallpox.
- Clustering - John Snow (Ucla, 2002, BBC, 2004). In his investigation of the cholera outbreak of 1854 Dr John Snow applied a meticulous process of interviewing to collect data. He used the information collected to develop a statistical map which clustered interview responses based upon the water pump which supplied water to the individual. This revealed that every victim had used a single supply of water and no non-sufferers had used that supply. Further investigation showed that this pump was in fact contaminated by a nearby cracked sewage pipe. This shows not only the power of the use of medical data for statistical purposes, but the benefits that can result from applying clustering techniques to that data.
- Association rules and classification - Henri Laboit (Pollard, 2006). Laboit extended the use of promethazine to treat mental disorders including schizophrenia by realising patterns in side effects from administering the drug during surgery on non mentally ill patients. This was achieved through identifying association rule style patterns to describe associations between focal and non focal attributes, for example combinations of relationships between diagnosis, treatment, symptoms, side effects and medications. Analytical techniques were employed to classify conditions exhibiting similar patterns of presentation and clinical testing was utilised to demonstrate the effect of applying an identified drug to control those classes of symptoms.

These techniques until recently were employed manually and hence were on a much smaller scale than we see today through the application of automated data mining systems, however they demonstrate the impressive potential for automated data analysis techniques to be applied with greater benefits and applicability than ever thought possible. There is a belief by some that the rate of medical breakthroughs of the calibre of those listed above has slowed dramatically since the 1970s (Horton, 2000). This could be attributed to the inability of the human mind to manage the volume of data available (Biedl et al., 2001, Lavrak et al., 2000) and that most if not all patterns in data which may reveal knowledge and which occur frequently enough to be noticed by the human analyst are now known. This adds significant weight to the argument for the application of more effective and efficient automated technologies to uncover the less visible knowledge or less frequent but equally important patterns in the data. We must however learn from history and ensure that the validation requirements for knowledge acquisition, as discussed previously, are adhered to by any automated process as for all other methods of knowledge acquisition even though this has been described as "the hardest part of the expert system development task" (Lavrak et al., 2000). Too often in recent work there has been a focus on developing new methods for determining the quality and value of outputs which does not take into consideration the many lessons we can learn from history, and is often little more than a process of reinventing an already rolling wheel (Ordonez et al., 2001).

3. The automation of medical knowledge acquisition

To automate it is not sufficient to simply understand how the process has occurred manually or how that process developed, although this is important in ensuring the results can be trusted. There is also a requirement to develop a seemless transfer from the clinical application of the process to the technical automated application of the same process and to provide accountability so that the results are trusted, justified and actionable in a clinical environment. In data mining systems, these accountability values are often measured through a concept termed 'interestingness' which essentially aims to measure the level of interest a user may have in the outputs provided by a system

3.1 What is interesting?

The term 'interest' is one which is widely accepted and used within data mining to denote output which has value on some level. This section is concerned with how we might define interest or value in clinical terms. The work discussed in this chapter began with the naive idea that interest can be measured and quantified for medicine as a homogenous entity as occurs in many other domains. This notion is now considered laughable at best. Medicine is not a homogenous entity and is not even a single entity in any contextual argument. Whilst it is defined generally (Oxford, 2007) as the science of studying, diagnosing, treating, or preventing disease and other damage to the body or mind, or treatment of illness by the ingestion of drugs, modification of diet or exercise, or other non-surgical means. The means by which those engaged in the practice of medicine achieve this outcome or engage in this activity varies widely depending upon individual needs, experience and data. If we are developing an automated system to assist or guide in this activity then the system must also be able to perform according to individual needs, experience and data. As discussed in the chapter introduction, data mining systems are able to manage a broad range of data types and analytical processes, but they are usually tailored to a user and hence are designed to

work with their individual technical proficiency and analytical requirements. Whilst work is ongoing in this area, uncertainty remains regarding the ability of automated systems to address the varied and fluid requirements of the user population. There are therefore two questions posed:

1. How can we identify the varied individual requirements of the user body?
2. Can we automate these requirements into a system which caters to a range of users?

To overcome this issue it is necessary to define what is interesting and what makes one thing more or less interesting than another and further to develop an algorithm to measure the extent to which the outputs of data mining conform to the user definition of interesting. As each of us finds different things interesting it is not possible to define a single specification for what is interesting to any group of people. Within the context of the focus of this chapter however, we can say that the degree to which something is deemed interesting can be quantified through the application of statistical methods. This method of value measurement is one which is applied in clinical testing where such measures as those shown in Table 1 are frequently used as a basis for determining which trial or trial arm has provided evidence that is clinically applicable or worthy of progression to the next level of testing (Gebeski & Keech 2003, Moser et al., 1999, Hilderman & Hamilton 2001, Geng & Hamilton 2006). Many of the same statistical methods are also applied within data mining systems but the provision of individual methods or combinations of methods are fixed and provide an inflexible analytical toolbox unlike in the clinical setting where any method can be chosen and applied. The requirement is therefore to provide a more flexible approach and not 'invent' another formula to determine the level of interest but to allow the user to determine how to define interest for each run or data set as occurs in clinical analysis. An added benefit in this is the ability to reinforce and support medical professionals control of their domain and the processes they apply to an analysis task. It should not be acceptable for technology to dictate how a medical professional (or any other) should practice.

Measure	Type	Application	Domain
P	Statistical	Determine degree of difference in results	Medicine
chi2	Statistical	Determine degree of difference in results	Medicine
chi2	Statistical	Result comparison	Medicine
Pearson's Correlation	Comparative	Measure of difference	Bio-inf
Euclidian Distance	Comparative	Measure of difference	Bio-inf
Cosine Similarity	Comparative	Measure of similarity of text	Linguistics
Support	Statistical	Probability, Frequency	Retail
Confidence	Statistical	Probability, Frequency	Retail
Accuracy	Domain	Determination of class membership	Medicine
Sensitivity	Domain	Measure of ability to find true positives	Medicine
Specificity	Domain	Measure of ability to reject true negatives	Medicine

Table 1. Some of the more commonly applied statistical tests.

In essence if the results of analysis are to be deemed interesting they must fall within defined thresholds as measured by one of more statistical method. The selection of appropriate methods is both objective and subjective. Objectively, certain qualities must be present in an interesting result and methods will be selected to provide evidence of this quality. For example if the result must be indicative of an accurate prediction of disease classification then such measures as sensitivity and specificity could be chosen as they are designed to measure this quality. Subjectively an analyst may place greater trust in one method over another due to experience or availability even though they both provide a method for measuring a particular quality. An example here may be the choice between using a p-value or chi^2 which can both provide similar quantified evidence. There is a long list of statistical methods which can be applied and they are selected based upon a combination of user objectivity and subjectivity which will potentially be different for each user. Although the provision of a fixed subset of available statistical methods represents the commonly applied data mining technique for evaluating interest, for the reasons discussed herein, it is not an appropriate methodology for the medical domain where there is no fixed notion of what is interesting or of how to measure the qualities which define interest. Whilst the process would be vastly simplified by ignoring the concept of individual interest it is necessary to overcome a number of issues with the application of data mining to medical data. The most important issue is that of non acceptence of the technology even though its benefits are many. Non acceptance is due to a defined set of factors:

- The complexity of medical data frequently results in a huge number of results; too many to be evaluated by the human user and a method for reducing the results to only those of greatest interest is necessary (Roddick et al., 2003; Ordonez et al., 2001)
- Each user has a trusted set of methods which are applied during clinical analysis and which are rarely seen in a system that is not purpose built for that user or their analytical requirements; this increases the cost of providing the technology, the frustration in trying to use the technology and allows the technology to dictate the analytical process rather than the other way around. Being able to facilitate and apply a range of interest definitions in a single system would open up the technology to a greater audience.
- Many users are not technically adept and do not trust something they do not understand; the provision of a recognised process which offers a personalised perspective within a generalised framework provides comfort and security.

To overcome these barriers to the technology it is neccesary to move away from a user focussed approach, which has in fact created many of the issues presented here, and towards an interest based approach which is guided by individual needs as it is the level of interest to each user that primarily determines the acceptability of results. If we can develop a generic approach to interest that can be individually adapted then we can apply this to develop data mining systems which can be similarly founded upon a generic principle that can be personalised for individual use. The provision of such a solution is the focus of the remainder of this chapter.

3.2 A role based approach

The issue of developing a flexible data mining system with the intention of enabling the generic production of results with an acceptable level of interest for each individual user has been the focus of recent work by the author. The approach has been to investigate the

concept of a role in various forms and its relationship to the concept of interest as defined above.

Each of us plays many roles in our daily lives as a student, doctor, nurse, teacher, parent, guitar player, amateur photographer etc. etc. Each of these roles will define a level of interest in the world around us, which will be determined by how relevant to each role the world is at any specific time. As each of us has a unique set of roles then each individual will have a corresponding unique set of interests or interest triggers, and different information will appeal to us at different times. Our role is therefore defined by how we measure the value of our interest in information that is presented to us. For example an Oncologist would most likely have been interested in an article on new cancer treatments in a Weekend Australian newspaper entitled "Hype or Hope?" (Cornwell, 2005), as it relates to their professional role and, even an evaluation of keywords contained in the article, would have revealed many matches to a similar evaluation of keywords in their set of interests. In contrast an Electrical Engineer would most likely not have had the same level of interest unless they also held the role of cancer sufferer or carer of a cancer sufferer for example. Therefore we can define a role as being a collection of quantifiable interests. The focus of work presented here has been to develop a system that will allow the identification of these interest sets and develop a method for determining how to measure and evaluate how strongly the information is able to trigger interest. To achieve this it was necessary to provide a quantitative evaluation of interest by evaluating the requirements of an interest set and measuring the applicability of the information or data mining results to that unique set.

3.2.1 The application of role

The application of a role in determining which data mining outputs are of relevance is more complex than simply looking for the presence of keywords. In the field of epidemiology for example the simple presence of the word 'flu' is not sufficient to trigger interest, there needs to be statistical augmentation to the information. In particular it needs to be shown that the incidence of the condition is sufficiently different to that expected for a population at a defined time. The difficulty is in determining which heuristics will give an acceptable measure by which we can include or exclude results for each role. A role based result evaluation engine has been designed in preference to other options including new heuristics and new heuristic combinations as these fixed solutions can not provide a generic answer to the issue of evaluating the level of interest for the health domain as a whole given the complexities noted above. It was necessary for an evolution in current thinking in the area and for single-user solutions to be discarded. A generalised solution to the issue of result reduction for this domain cannot be achieved due to the broad range of roles and requirements to be addressed. Early work has since evolved into developing a system that can incorporate the range of roles without the need for a separate system for each. As it is the role that determines how the strength of general interest is measured, it was a natural step to discriminate the analysis by the role of the current user as defined by a set of measurable interests. Users should be able to analyse and focus their data mining outputs using a single system regardless of their speciality or analytical requirements. Whilst the needs of each role are unique there is also considerable overlap and the heuristics required to determine interest strength varies from role to role and also within each user role depending upon the nature of the analysis being undertaken (Kuonen, 2003; Bresnahan,

1997; Tan et al., 2002). A system with a high level of flexibility and methods to facilitate user definition was deemed to provide the best use of resources to accommodate this.

Role based access models have been successfully implemented in a wide range of domains and have demonstrated an ability to overcome issues such as those seen in health including a need for careful management of sensitive information and the need to provide enterprise level security policies which discriminate on a local level based on the role of the user (Cabri et al., 2004; Ferrailo & Kuhn, 1995). Role based access models have provided a fixed framework from which to apply highly flexible system definition and this concept was the major attraction in the creation of a role based results evaluation process for data mining applications in the health domain.

The following features of role based systems have been adapted and incorporated into a hypothesis engine as discussed in Section 4.

- The accommodation of roles that allow for overlapping requirements and measures.
- The ability for a user to have more than one role at a time.
- The ability to enforce constraints on data access where required to accommodate ethical sensitivities.
- The ease of modifying the role of a group of users to accommodate new technologies or methodologies.
- The ability to constrain at a global level and provide flexibility at a local level.

Whilst the choice of interest strength heuristics will often fluctuate little across mining runs, some vary greatly, become redundant or require supplementation by new or existing measures. This high level of flexibility is not currently available in documented data mining systems. Algorithms and selected heuristics are applied singly or in a fixed combination with others within a specific system designed for a specific use. The ability to utilise the interest role as a means of selecting and applying a significant number of the range of measures in a unique combination as required has not been documented and is believed to be a novel approach to the issues presented here. Support for such an approach has been provided by health domain professionals and domain based publications including the Medical Journal of Australia (MJA) which stated that appropriate statistical methods for analysing trial data are critical and suggested that the statistical methods used in each trial should be specifically tailored to each analysis (Gebski & Keech, 2003). Each specialist field and role has its own requirements and hence the level of flexibility in data mining software packages must be equally flexible, open to adaptation and tailored for the user role at run time.

As discussed in the previous section, there are many methods which can be applied to measure the strength of interest a user may feel towards any information. An initial aim was to group these measures based on the role that uses them, this was rejected due to the overlaps discussed earlier. The aim was thus modified to group the measures into classes based upon their type and the characteristic of interest they are able to quantify, thus allowing each role to select and apply them as required. While there is no fixed notion of what defines interest strength for a particular role in each instance, there is agreement on the characteristics that indicate strength and these can be grouped and measured to quantify their level of expression in results presented. These classes were verified in discussion with a range of medical professionals during a work in progress seminar (Workshop in population health, 2005) presented to by the author. In attendance were medical specialists from several fields including cardiology, epidemiology, biostatistics, nursing, clinical

research and government and all agreed that the proposed classes defined the qualities they looked for during hypothesis development and results testing. It was noted that whilst most of those present could not adequately describe what determined a strong interest in a result for the domain generally, it was felt that the classes presented would provide an acceptable quantification for any role within the domain if applied uniquely for each role or field. It was also proffered that each test result was considered individually depending upon their needs at the time often the heuristics employed were often not selected until the time of evaluation thus strengthening the argument away from a generalised approach. This, in fact, emphasised the need to utilise traditional measures but in a flexible combination for each evaluation.

By allowing a subjective selection of heuristics and evaluating their application objectively it is possible to take the outputs of data mining and measure their value uniquely and flexibly for each role rather than utilising a unique but fixed sub set of heuristics for each system. Based upon the values achieved by each heuristic, unqualified mining outputs can be eliminated from presentation thus providing only those outputs which meet the requirements of the role and adequately contain the desired characteristics to be of interest.

Six classes or criteria are provided for interest strength measurement and each of these may require a number of statistical, comparative or other tests to determine the overall strength for each criterion. The individual values are then combined to provide a comprehensive measure of interest strength for a mining output based on the total requirements for the user role. A greater strength suggests an output that is more likely to be of value to the role that defined the heuristics and their scopes. The criteria for measuring interest and hence strength of new information or knowledge patterns produced through data mining are discussed following.

- Novelty - Is it unknown in the body of domain knowledge? This is more complex than simply not duplicating existing knowledge or presenting expected patterns. New patterns based upon existing knowledge may still be of interest if the strength or content of the new knowledge differs sufficiently from that which is expected. For example, medical professionals would reject as new a pattern which states that 3.6% of pregnant women develop gestational diabetes mellitus (GDM) as this is known and expected knowledge even though it would have sufficient strength by some traditional measures to warrant further investigation (Stone et al., 2002). However if a pattern were to report that the incidence rate of GDM in a data set primarily for a North Asian population was 3.6% then the interest in this may be greater as the rate would be expected to be higher. Hence it is the pattern novelty as a whole which is being evaluated and which thus determines the strength of interest. There are a number of measures that can be applied to quantify the expectedness or similarity of hypotheses to existing knowledge and it may be necessary to test this criterion using several classes of tests to adequately assess the novelty of a pattern.
- Applicability – Is it relevant to the current user? This infers that either some contextual information is required, or that previous patterns are tagged as interesting (or not) so that the system can learn and reference. The definition of applicability (or relevance) is context based and should be maintained on an individual level. An outlier that is strong in every aspect except for prevalence may not be relevant to an epidemiologist as it is not representative of the population but still may potentially be of interest to a clinical specialist or medical researcher and should be tagged for reference by that role. The implication is that any derived pattern produced from a medical data store is

potentially valuable to some role in the medical domain. If accepted, this suggests the importance of strength determination at a role based level to ensure that each role sees only patterns they are most likely to have an interest in and be able to act upon but that no strong pattern is omitted completely from consideration.

- Relativity – Is it valid relative to the data from which it originated or a class of object that it describes? Once again the applicability of this criterion is determined by the context within which it is measured. Within epidemiology it is important for pattern to be shown to be applicable for a generalised population. Results therefore need be demonstrated to apply across the human race or a definable sub section of it. A recent study published in the MJA discussed a potential but low correlation between passive smoking and breast cancer (Smith & Ebrahim, 2004). Whilst the link was biologically plausible in 1999 it was not deemed to be representative of the female population in an epidemiological sense and hence was not deemed interesting. Further work was done which focussed on the effect of environmental tobacco smoke across the age variable specifically. It is now accepted that there is enough evidence to suggest that passive smoking specifically in the early years of a females' life has a measurable impact upon the incidence of breast cancer later in life. Investigation at a finer granularity resulted in a hypothesis that is accepted as representative of a defined sub section of the population. This suggests that strength should be measured for all applicable classes, not only the most obvious or highest ranking. Patterns also need to be shown to be representative of the data set from which it came and there are standardised checklists such as that provided by CONSORT (Consolidated Standards of Reporting Trials) (Lord et al., 2004, Gebeski & Keech 2003, Altman et al., 2001) which are widely used within medical research and should be incorporated into the planning of data mining systems.
- Provability – Can it be proven through clinical testing? This reflects the actionability of the outcomes of data mining and incorporates the need to adhere to guidelines such as CONSORT discussed earlier. Whilst there are perceived difficulties in automatically determining what could be tested clinically, there are several requirements which define what the foundations of a clinical hypothesis should be and these should be present in hypotheses in the form of patterns derived through data mining also (Lord et al., 2004) . For organisations that adhere to research guidelines, it is important that the pre-requisites are met for further work so that the potential for follow up clinical testing is not prevented. This criterion aims to ensure that potential hypotheses are not rendered inactionable due to the methodology employed for their derivation rather than trying to determine what will be actionable.
- Understandability – Can it be understood through appropriate presentation? New knowledge that cannot be described easily or accurately is of little use. The inability for the human brain to assimilate and perform functions upon large amounts of complex data is the very foundation upon which the field of data mining was based. When presenting patterns, this must be given due consideration. An overly complex or lengthy pattern may be overlooked in favour of those that can be read and understood quickly. Consideration must also be given to domain specific terminology and semantic hierarchies (Ashby & Smith, 2002). This will ensure that patterns are presented using uniform, accurate and appropriate terminology (Bojarczuk et al., 2001, Lavrak et al., 2000). Results should also be presented via a medium that is accepted as standard by each role or domain and there is a body of work in the fields of visualisation and linguistics that is attempting to address some of these issues.