Establishing a Lineage for Medical Knowledge Discovery

Anna Shillabeer[1,2] and John F. Roddick[1]

[1] School of Informatics and Engineering,
Flinders University,
PO Box 2100, Adelaide,
South Australia 5001
Email: {anna.shillabeer, roddick}@infoeng.flinders.edu.au

[2] Heinz School Australia
Carnegie Mellon University
Torrens Building
220 Victoria Square
Adelaide SA 5000

Abstract

Medical science has a long history characterised by incidents of extraordinary insights that have resulted in a paradigm shift in the methodologies and approaches used and have moved the discipline forward. While knowledge discovery has much to offer medicine, it cannot be done in ignorance of either this history or the norms of modern medical investigation. This paper explores the lineage of medical knowledge acquisition and discusses the adverse perceptions that data mining techniques will have to surmount to gain acceptance.

Keywords: Medical and Health Data Mining.

1 Introduction

The nature of data mining research is that it requires a second discipline to be validated as useful. To do this, data mining must adapt to this second discipline and conform to the norms expected of that discipline. In contrast to many disciplines where data mining has been applied, medicine has a strong, established and accepted research methodology and application of data mining technology that falls outside this, however well-meaning, will struggle to be taken seriously.

This paper thus explores the history of knowledge acquisition in medicine and extracts from this history some important issues that data miners should take into account when mining medical data.

The paper is organised as follows. The next section explores the history of knowledge acquisition methodologies in medicine, while Section 3 discusses the role of intuition and serendipity in many medical advances. Section 4 then briefly discusses the role of data mining to the medical context and includes some examples of where data mining techniques have implicitly been used. Section 5 then outlines a number of arguments that have been raised against the adoption of data mining in medicine and briefly discusses each in turn.

2 Knowledge Acquisition in Medicine

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 were 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 people of the value of long held medical beliefs. He 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 process has survived to the present day and is reflected not only in the provision and acceptance of new knowledge but also in 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 contexts are discussed here briefly to demonstrate the extent and foundations of medical knowledge acquisition debate in the ancient world.

2.1 The Graeco-Roman Context

- Methodists

  The first prominent physician practising the Methodist philosophy was Hippocrates of Cos (460-380 B.C.) – the so-called *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 which it contains. The body of work was, how-

ever, 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 founded their knowledge on an understanding of the nature of bodily fluids and developing 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 1977). 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 prominent in this group (Cosans 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 were not deemed sufficient (Brieger 1977). 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 the cause and effect of physical change in the body (Horton 2000).

- Empiricists

  The Empiricists believed that it was not sufficient 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, better known and more frequently quoted empiricists (Brieger 1977). He was particularly interested in testing the theories proposed in the Hippocratic Corpus, especially given its frequent contradictions. His work was also produced 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 The Islamic Context

- The Methodists (Ashab al-Hiyal).

  In direct parallel to the Methodist philosophy discussed earlier, this sect 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 n.d.).

- The Empiricists (Ashab al-Tajarib).

  Islamic 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 (Muhaqqiq n.d.):

  - Incident (*ittifaq*) – 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 (*iradah*) – denotes an event experienced by choice, such as taking a cool bath to reduce a fever.
  - Comparison (*tashbih*) – a technique employed by a practitioner whereby it is noted that a technique 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.
  - The adoption of a treatment that was used in another similar case (*naql min shay' iki shabihihi*) – a technique whereby the physician applies a treatment for a similar presentation for a presentation which has not been encountered before. For example, 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 the 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 might have been treated differently 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 scientific belief and knowledge should be derived from experience and observation, tempered by 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 to correctly diagnose and treat any condition. They define changes as being of two types (Muhaqqiq n.d.):

  - *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 well-being.

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 the Graeco-Roman Rationalists.

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 n.d.).

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 a quorum of medical knowledge. In the 11th century a Spanish scholar, Said Al-Andalusi, stated that he believed that the Indian people *are 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 Andalusi 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 to acquisition of knowledge regarding the human state was the Upanishads which were believed to have been written between 1500 and 600 B.C.E. and were concerned with knowledge regarding the spirit, soul and god (South Asian History Project 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. This formed the basis for the emergence of the rationalist period. Early rationalists included the Lokyata, Vaisheshika and Nyaya schools. 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* (South Asian History Project 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 Vaisheshika school's main achievement in the progression of human knowledge was in their development of a process for the classification of entities in the natural world, and in their hypothesis that all matter is composed of vary small particles with differing characteristics. (South Asian History Project 2002). Their theory stated that particles, when combined, gave rise to the wide variety of compounds found upon the earth and allowed them to be classified by 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 (South Asian History Project 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 to provide apparent evidence for theories or knowledge (South Asian History Project 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 and generally composed of specious arguments; and *chal*, the use of language to confuse or divert the argument.

Further to this again a set of five 'logical fallacies' was 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 and many remain applicable in modern scientific research.

- The Jains are worthy of note not because of the size of their impact on the process of acquisition of scientific knowledge but due to their identification of a truth matrix which demonstrated that there are more possible outcomes from scientific research than simply true or false as shown in Table 1.

	Proved	Indeterminate
True	⋆	
False	⋆	
True or false	⋆	
Indeterminate		⋆
True or indeterminate		⋆
False or indeterminate		⋆
True or false or indeterminate		⋆

Table 1: The 7 states of truth according to the Jains

Prior to the work of the Jains, scientists described their outcomes only in terms of true or false and did not consider that there may be degrees of truth or that a hypothesis might not have been proved or disproved but may remain open to debate or require further testing to gain a conclusive result.

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 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 foci remain the topic of debate 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 rightly informs the focus of much research.

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. Indeed, 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 and often untrusted application of scientific processes. Unfortunately, this often 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 chance, tenaciousness and/or the application of non-conventional methods to obtain evidence.

- James Lind (1716-1794) (Buck et al. 1988). Based upon an unsubstantiated personal belief that diet played a role in the development of scurvy on naval vessels, Lind performed limited randomised trials to provide proof and then published his *Treatise on the Scurvy*, which is still relevant to this day.

- Edward Jenner (1749-1823) (Sprang 2002). During his apprenticeship, Jenner overheard a milkmaid suggest that those who have had cowpox did 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.

- John Snow (1813-1854) (Burke 1985). 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.

- Alexander Fleming (1881-1955) (Mulcahy 1996). 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.

- Carlos J. Finlay (1833-1915) (Adams 1992). Finlay's observations regarding cholera, although similar to Snow's, were not taken seriously because of a perceived criticism of the local authorities. His observations regarding the mosquito as a vector in the transmission of Yellow Fever were also nearly dismissed and it was 20 years before his theory was taken seriously.

- Henri Laborit (1914-1995) (Pollard 2006). During his ward visits, 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.

- Robert Edwards and Patrick Steptoe (1925-, 1913-1988) (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.

- Barry J. Marshall (1951-) (Marshall 1998). 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 depart from conventional methodologies to facilitate the acquisition of knowledge, although there is always a requirement to subsequently provide substantiated proof and an argument based upon accepted scientific principles.

The applicability of this notion of departing from conventional methodologies is particularly relevant to data mining research with its focus on the application of new techniques and technologies which already have demonstrated an ability to provide an important impetus to the acquisition of knowledge in other domains. 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.

Despite the methodology for producing the evidence required for knowledge acquisition, the above examples all fulfilled a number of basic requirements prior to acceptance. These requirements are summarised below:

1. Replication of results. For data mining, this means that datasets must be, at least in theory, publicly available. Moreover, as Freitas (2000) points out, association rule mining generates the same set of rules for every subsequent run over the same data, whereas some classifiers can be unstable, generating markedly different classifiers for small changes to the input dataset.

2. Non contradictory results. There are many data mining algorithms that will produce results that include an apparent contradiction but few that attempt to detect these contradictions and explain them.

3. Scientifically justified theories and hypotheses.

4. Ethical methodologies and measures.

5. Results demonstrated to be representative of the population. This means that accepted methods of statistical confidence must be adopted.

6. Results derived from sufficient numbers of cases. Data mining works best over large quantities of data. Running data mining over small datasets is unsound and, arguably, not within the scope of data mining technologies.

7. Publicly documented processes and results.

Some of the impediments to the adoption of data mining in medicine are discussed in Section 5 and demonstrate the need to understand and apply these requirements.

4 Medical knowledge acquisition through data mining

One of the more commonly quoted definitions of data mining is that it is a *non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data* (Houston et al. 1999). Data mining is essentially part of a larger knowledge discovery process whereby data is transformed into information and knowledge is then extracted from that information with the focus being on the identification of understandable patterns in data. This definition is used as a foundation for the development of many medical data mining systems. However, data mining has the potential in the medical domain to expand this definition even though use to date has been sparse (Forsstrom & Rigby 1998).

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 has been the trend in other domains. Very little health data mining is purely exploratory and hence it is generally not applied to provide novel knowledge, that is, to identify new 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 due to missing attributes leading to incomplete information, may not be statistically valid by trusted measures or may simply not relate to the specialisation of the user and is therefore irrelevant. Hence data mining systems to date have generally been developed for a specific user or data type and for a specific purpose.

The health domain is complex and standardised data mining techniques are often not applicable (Cios & Moore 2002), however, there are four procedures that are frequently documented.

1. Production of association rules;

2. Clustering;

3. Trend or temporal pattern analysis; and

4. Classification.

These processes are applied to provide three generalised functions:

1. Prediction;

2. Expert decision support;

3. Diagnosis.

It should be noted that 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.

- Data sampling - James Lind (Katch 1997).

  Lind performed small randomised trials to provide proof of the cause of scurvy. In his position as Naval doctor he could 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 on 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 mining - 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 smallpox 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 (Burke 1985).

  In his investigation of the cholera outbreak of 1854, 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 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 previously thought possible.

5 Perceived Impediments to the Adoption of Data Mining in Medicine

The history of medical knowledge acquisition can be seen to inform some of the criticisms of medical data mining which have led, in some cases, to the technology being overlooked as a tool. In particular, the (initially at least) exploratory nature of data mining seems to negate the need for a hypothesis. Such criticisms must be addressed and this section discusses six of the strongest arguments cited against the application of data mining in medicine.

Data mining outcomes are seen as generalisations and not verified for medical validity or accuracy. (Elwood & Burton 2004, Milloy 1995)

Medicine is a highly complex domain for which data mining processes were not designed. Frequently they originated in response to changes in commerce or management practices where there was no methodological need to substantiate results on the basis of protocols or domain knowledge. Medicine has requirements which are outside the original scope of the technology, and to be applicable to a science which is concerned with critical decision making there is a need to modify the technology to reflect this different environment.

Whilst this is a serious issue, it is often borne from misrepresentation of the results of data mining rather than from the process itself. There is a need for careful consideration of the language used when reporting results (Maindonald 1998). It is possible for the results to be specific but for the language of reporting to generalise the message. For example, Elwood & Burton (2004) describe a case where a mining outcome showed that smoking does not have a direct link with skin cancer. However the resulting media story reported that smoking is not linked with cancer generally. While a scientific data mining process was applied the language of information presentation was misleading and the resultant reporting was inaccurate and medically invalid. Medicine is especially sensitive to this form of information distortion and the consequences have the potential to be life threatening, politically sensitive, costly and persistent which is rarely the case in other domains.

There is little to sustain this argument in light of recent work in the field. By the application of suitable statistical methods, evaluation of all results and applying industry accepted standards there is no reason to believe that data mining cannot provide effective validation and accuracy checking processes (Gebski & Keech 2003, Shillabeer & Roddick 2006). Three steps have been suggested to safeguard against this criticism (Smith & Ebrahim 2002).

1. Results should not be published on the basis of correlation alone.
2. An explanation should be provided with the results to provide clarification e.g. A definition of the unique quality of the allergen that triggers the alleged immune response.
3. Results should be replicated, confirmed and documented prior to publication.

These steps are not part of standard data mining methodologies but are required to be undertaken if the mining of medical data is to overcome criticism, be viewed as 'good science' and gain trust in the medical community.

Associations are not representative of other similar attributes and do not consider other potential contributors. (Milloy 1995, Smith & Ebrahim 2002)

In a medical context, relationships found between one allergen and symptoms must be substantiated through analysis of similar allergens or the same allergen in other temporal, spacial or demographic instances. If this cannot be shown it suggests that there is not a conclusive argument for cause and effect or that some other catalyst or cause has been missed (Smith & Ebrahim 2002). Again, data mining was not designed to do this however this should not be a preventative. Methods are available to achieve this where it is important to determine the semantic closeness of results. Criticism often focuses on data dredgers who promote results as facts rather than being indicative of a possible scenario requiring further investigation. Where an association is found it is important to compare this with other associations or to apply a clustering algorithm to group semantically and determine where there is similarity or otherwise to other attributes or rules.

P-values are set arbitrarily and therefore the results cannot be trusted. (Milloy 1995, Smith & Ebrahim 2002)

P-values may be applied in two ways: to evaluate and discriminate the acceptability of mining results and, as a guideline or tool for reducing irrelevant outcomes. Data mining can also be applied in divergent modes; to show what the common patterns in data are, or to show where common patterns are refuted in the data. Obviously the p-values required for these two mining runs would be different thus obviating the need for a range of p values to be applied. An issue arises where the p-value is modified iteratively until the outcomes meet some predefined need. It is important to always set heuristic thresholds in context of the specific analysis being done and in fact a calculation for p should be only one test among many (Gebski & Keech 2003, Shillabeer & Roddick 2006). In the medical domain, attribute-value relationships which occur frequently, and hence have a low p-value, are likely to be known already and would generally be of little if any interest. This is a major difference between traditional data mining applications, where generally the events which occur most frequently are of the greatest interest and hence have a similar p threshold, and applications in the medical domain where frequency is not a conclusive determinant in defining the usefulness, validity or applicability of results and hence may require varying p values.

Associations between attributes are dependent upon the data set being analysed and

are not representative. (Smith & Ebrahim 2002)

There is often a poor approach to the collection and description of data sources and samples which is not consistent with the process of data mining and other scientific methodologies (Milloy 1995, Maindonald 1998). For results to be accepted the data source should be from an identifiable population with defined characteristics such as location, demographics, and proportions (Smith & Ebrahim 2002). In a clinical research setting this is overcome using protocols and guidelines to ensure that results are representative and can be replicated. One such protocol is CONSORT which is used globally by medical researchers and is endorsed by a number of prominent journals (Moher 1998).

In epidemiological studies, this problem is exacerbated by the use of external, non medical domain specific datasets that, while generally accurate and reliable, where not collected for the purposes of data mining (Roddick et al. 2003).

Data mining provides validation through tools such as artificial intelligence and neural nets being applied in the knowledge mining step to sample the data, provide outcomes then automatically test them on the whole data source to show that the outcome holds true for all available data not just one small subset (Smith & Ebrahim 2002). Data mining is a highly intensive machine process which utilises huge processing power, memory and time. Data sampling is often used as an initial step to reduce these constraints but correct utilisation may help to overcome this criticism also.

Data mining is simply a desperate search for something interesting without knowing what to look for. (Milloy 1995, Smith & Ebrahim 2002)

Exploratory mining, which is not constrained by user expectations, can uncover unexpected or unknown knowledge with wide reaching benefit and can be utilised to review and extend current medical knowledge. With the wealth of data being produced daily in the medical field the argument that it should not be used in an exploratory fashion to at least note important changes in data patterns demonstrates a misunderstanding of the potential value held therein. It is argued by some (Maindonald 1998, Smith & Ebrahim 2002) that it can be beneficial to look simply for something interesting rather than make an assumption about what is present in the data as if we only ever look for what is known we will potentially never find anything new and progress cannot be made. Provided this is a result of a scientific process then further mining or clinical trials can be undertaken for evidence to substantiate the initial findings. This criticism is only valid where the search is for anything interesting even if only minimally and where there is little or no validation.

Data mining displaces research and testing and presents results as facts requiring no further justification. (Milloy 1995)

Contrary to the criticism, data mining in medicine is generally viewed as an efficient tool for enhancing the work done in the field rather than as a replacement for it (Maindonald 1998). Its value is seen as a process of *automated serendipity* that stimulates and supports testing

rather than replaces it. When considering the use of mining outcomes there are two questions often asked; is this result representative of what has been recorded over time?, and can the analysis outcome be verified through real world application? (Smith & Ebrahim 2002). Whilst the first can be answered with some conviction by data mining the second requires clinical input and hence the process of providing trusted knowledge from data requires a collaborative effort by automated and clinical processes. When we consider that time from hypothesis to application of new knowledge is often measured in decades we should feel compelled to find new knowledge as quickly as possible and data mining offers the ideal tool for this.

6 Conclusions

History and current practice are important issues and need to be taken into account when constructing or justifying data mining techniques in medicine. This paper has outlined some of these issues.

There is a belief by some that the rate of medical breakthroughs of the calibre of those listed above has slowed dramatically since the 1970's (Horton 2000) This could be attributed to the inability of the human mind to manage the volume of data available 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 above, are adhered to by any automated process as for other methods of knowledge acquisition.

References

Adams, J. R. (1992), *Insect Potpourri: Adventures in Entomology*, Sandhill Crane Press.

al Andalusi, S. (1991), *Science and the Medieval World: "Book of the Categories of Nations"*, University of Texas, Austin.

Brieger, G. H. (1977), 'H l coulter. divided legacy: A history of the schism in medical thought.', *Isis* **69**(1), 103–105.

Buck, C., Llopis, A., Nájera, E. & Terris, M. (1988), *The Challenge of Epidemiology: Issues and Selected Readings*, World Health Organization.

Burke, J. (1985), *The Day the Universe Changed*, BBC, London.

Cios, K. & Moore, G. (2002), 'Uniqueness of medical data mining', *Artificial Intelligence in Medicine* **26**(1-2), 1–24.

Cosans, C. E. (1997), 'Galen's critique of rationalist and empiricist anatomy', *Journal of the History of Biology* **30**, 35–54.

Elwood, J. & Burton, R. (2004), 'Passive smoking and breast cancer: is the evidence for cause now convincing?', *Medical Journal of Australia* **181**(5), 236–237.

Fauser, B. C. & Edwards, R. G. (2005), 'The early days of IVF', *Human Reproduction Update* **11**(5), 437–438.

Forsstrom, J. J. & Rigby, M. (1998), Addressing the quality of the it tool - assessing the quality of medical software and information services, Technical report, University of Turku and Keele University.

Freitas, A. A. (2000), 'Understanding the crucial differences between classification and discovery of association rules: a position paper', *ACM SIGKDD Explorations Newsletter* **2**(1), 65–69.

Gebski, V. & Keech, A. (2003), 'Statistical methods in clinical trials', *Medical Journal of Australia* **178**(4), 182–184.

Gopal, K. (1997), Rationalist victory, Technical report, Online at http://www.iheu.org/node/668.

Hanson, A. E. (2006), Hippocrates: The "greek miracle" in medicine, Technical report, Online at www.medicineantiqua.org.uk/sa_hippint.html.

Horton, R. (2000), 'How sick is modern medicine?', *The New York Review of Books* **47**(17).

Houston, A., Chen, H., Hubbard, S. M., Schatz, B. R., Ng, T. D., Sewell, R. R. & Tolle, K. M. (1999), 'Medical data mining on the internet: Research on a cancer information system', *Artificial Intelligence Review, special issue on the Application of Data Mining* **13**(5-6), 437–466.

Katch, F. I. (1997), History makers, Technical report, Online at www.sportsci.org/news/history/lind/lind_sp.html.

Kaul, M. & Thadani, S. (2000), Development of philosophical thought and scientific method in ancient India, Technical report, Online at http://members.tripod.com/ INDIA_RESOURCE/scienceh.htm.

Maindonald, J. (1998), New approaches to using scientific data- statistics, data mining and related technologies in research and research training, Occasional paper, Australian National University.

Marshall, B. J. (1998), Peptic ulcers, stomach cancer and the bacteria which are responsible, Technical report.

Milloy, S. (1995), *Science without sense. The risky business of public health*, Cato Institute, Washington DC.

Mohaghegh, M. (1988), 'Miftah al-tibb wa minhaj al-tullab (a summary translation)', *Medical Journal of the Islamic Republic of Iran* **2**(1), 61–63.

Moher, D. (1998), 'CONSORT: An evolving tool to help improve the quality of reports of randomized controlled trials', *Journal of the American Medical Association* **279**(18), 1489–1491.

Muhaqqiq, M. (n.d.), 'Medical sects in islam', *al-Tawhid Islamic Journal* **8**(2).

Mulcahy, R. (1996), *Diseases: Finding the Cure*, The Oliver Press, Inc.

Pearcy, L. (1985), 'Galen: a biographical sketch', *Archaeology* **38**(6 (Nov/Dec)), 33–39.

Pollard, R. (2006), Fortuitous discovery led to a revolution in treatment, Technical report, Online at http://www.smh.com.au/news/science /fortuitous-discovery-led-to-a-revolution-in-treatment/ 2006/10/11/1160246197925.html.

Roddick, J. F., Fule, P. & Graco, W. J. (2003), 'Exploratory medical knowledge discovery : Experiences and issues', *SigKDD Explorations* **5**(1), 94–99.

Shillabeer, A. & Roddick, J. F. (2006), 'Towards role-based hypothesis evaluation for health data mining', *Electronic Journal of Health Informatics* **1**(1), e6.

Smith, G. D. & Ebrahim, S. (2002), 'Data dredging, bias or confounding', *British Medical Journal* **325**(21-28 December 2002), 1437–1438.

South Asian History Project (2002), Philosophical development from Upanishadic metaphysics to scientific realism, Technical report, Online at http://india_resource.tripod.com/upanishad.html.

Sprang, K. (2002), Dr. Edward Jenner and the smallpox vaccination, Technical report, Online at http://scsc.essortment.com/edwardjennersm_rmfk.htm.