Exploratory Medical Knowledge Discovery : Experiences and Issues

John F. Roddick Peter Fule
School of Informatics and Engineering
Flinders University of South Australia
GPO Box 2100, Adelaide 5001
South Australia
{roddick,peterf}@infoeng.flinders.edu.au

Warwick J. Graco
Program Review Division
Health Insurance Commission
Canberra, ACT 2600
Australia
w.graco@hic.gov.au

ABSTRACT

The application of data mining and knowledge discovery techniques to medical and health datasets is a rewarding but highly challenging area. Not only are the datasets large, complex, heterogeneous, hierarchical, time-varying and of varying quality but there exists a substantial medical knowledge base which demands a robust collaboration between the data miner and the health professional(s) if useful information is to be extracted.

This paper presents the experiences of the authors and others in applying exploratory data mining techniques to medical, health and clinical data. In so doing, it elicits a number of general issues and provides pointers to possible areas of future research in data mining and knowledge discovery more broadly.

Keywords

Exploratory Data Mining from Health and Medical Data, Clinical Knowledge Discovery.

1. HISTORY AND MOTIVATION

Data mining algorithms and knowledge discovery frameworks have been successfully applied in a number of application domains including telecommunications, commerce, astronomy, geological survey and security (q.v. [16; 20; 34] and many others). For a number of reasons, the application of the technology to large medical datasets has necessarily been more circumspect. However, while data mining from medical, health and clinical data[1] can be problematic, it promises substantial rewards and has attracted some interest (q.v. [3; 21; 25; 26; 27; 29; 30; 31; 35; 44; 47] and particularly [9]). Much of this work reports on specific examples of applying explanatory mining techniques to particular, narrowly defined medical problems. Very little of the research is exploratory and, for some good reasons, very few broad exploratory analyses have been performed and reported.

The history of medical science as a discipline is extensive, sometimes tragic, certainly heroic and often serendipitous. On the last point, there have been a number of instances in which the detection of trends and anomalies across populations have resulted in significant advances – John Snow's observations regarding the transmission of cholera in the 1850s; Edward Jenner's observations regarding the connection between cowpox as a degraded form of smallpox and the ability for cowpox to provide an immunity; the statistical connection between schizophrenia rates and a person's month of birth; through to recent observations regarding deaths by drug overdose and the time of the month – the list is extensive.

There are a number of issues that must be addressed before sensible data mining can occur. For example, the available datasets range from those that are convenient, accurate, indexed and well-managed to those that are incomplete, inconsistent, potentially inaccurate and extremely large. Moreover, unlike other (arguably simpler) domains, the medical discipline itself is diverse, complex and, to an outsider, relatively opaque. It thus requires an active collaboration between the domain specialist and the data miner.

In addition, the fragmented and distributed nature of health data between GP surgeries, hospitals (both public and private), insurance companies and government departments poses substantial challenges for data integration and therefore for data mining in terms of the trust that can be placed in a result and the semantics of a derived rule. For example, in Australia, Health Insurance Commission (HIC)[2] data does not include the patient diagnosis for each episode and where the data is used for research purposes the diagnosis sometimes has to be inferred from the pathology tests carried out or from the medications prescribed.

In short, using medical data pushes current mining technology and processes to their limits and it is this aspect that promises to provide a practical insight into some of the possible future directions for knowledge discovery systems more generally. This brief paper reports on some issues and the experiences to date, both by the authors and by others (as reported in the literature), in applying data mining techniques to medical data.

[1]The terms *medical*, *health* and *clinical* are not synonymous but for brevity in this paper we will use the term *medical* to include all three.

[2]HIC is the Australian Government statutory authority established in 1974 to administer Medibank (a universal health insurance scheme) and later Medicare (Australia's provider of universal health care services).

2. ISSUES IN MINING MEDICAL DATA

In practice, there are two forms of data mining applied over medical data. Explanatory or confirmative data mining has been applied fairly widely to determine, for example, the optimum decisions to take under various clinical conditions or to predict an eventual insurance claim amount based on early diagnoses and the circumstances known at that point. On the other hand exploratory data mining has, to date, been applied relatively rarely.

In this section we discuss a few of the major issues which complicate exploratory medical data mining in relation to the various other domains in which it has been found to be useful. It is acknowledged that some of these issues *do* occur in other domains although rarely with the same level of interplay and difficulty.

2.1 The Investigative Method

While much useful medical data mining can be undertaken using more *conventional* mining methods, (both those used pragmatically for quick inspections and those facilitated by frameworks such as CRISP-DM [7]) other approaches often need to be considered. For example, the dominant medical research paradigm, the null hypothesis experimental method, used in medical / bioscientific research differs from that adopted by many other fields and often requires a change

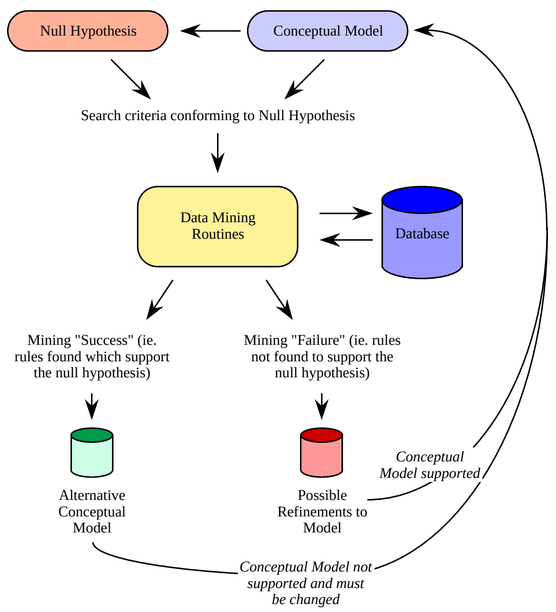

Figure 1: Scientific Knowledge Discovery Framework after [37].

to the knowledge discovery framework being employed. For example, we might look for an absence of conflicting data or we might generate a set of expected rules that are then compared with a set generated from the null hypothesis and for which we are searching for a lack of intersection between them. The framework in Figure 1 shows schematically the reversal of the process needed for this domain. To our knowledge, little work has been undertaken in this regard.

2.2 Longitudinal, Temporal and Spatial Support

The incidence of disease and their methods of remedy are closely linked to their order and their temporal and spatial occurrence relative to other episodes, intervals or locations. The use of static techniques thus oversimplifies (or may hide [33]) possible relationships and thus support for longitudinal, temporal and spatial semantics within the mining process is highly desirable. Episodic data is often the key to good data mining – it is a commonly held view that to find out what is really happening in health we need to focus on patient episodes.

Expanding the mining activity to include time and space multiplies the options available to the miner and the technique used needs to relate to the questions that are commonly asked [38]. Such questions include *What changed (or didn't change) when ...*, *What are the differences between ...*, *What values/associations changed with ...* or *What are the changes in ....* Note that the *change* may relate to a function of the data values rather than the data values *per se*. Some of the techniques that have been found to be useful include:

- **Anomaly Detection** – discovering changes in either item values or the strength of rules between items that differ markedly from either expected patterns or other values across that time period. This has some similarities to work done in intrusion detection (q.v. [24]) but to date little work has been undertaken using medical data.

- **Difference Detection** – discovering differences between rulesets that should exhibit similar properties. Comparing statistically significant differences between rules generated from comparable datasets from different hospitals, for example, can discover differences in the way in which each location operates. Work in other domains (such as signal processing) as well as recent work in inter-transaction rule mining [46] may be able to be modified for this purpose.

- **Longitudinal X Analysis** –this technique looks across a series of X, where X can be itemsets, rulesets, clusters, characterisations, etc., and characterises the (interesting) changes to them (cf. [41; 46]). For example, *Longitudinal Itemset Analysis* can detect patterns in the strength of associations between values and can thus answer questions such as *Find all itemsets for which the support is monotonically increasing* or *Find other associations that vary in line with itemset Y*. Similarly, using *Longitudinal Cluster Analysis* we can detect changes in the number or cohesion of clusters.

- **Temporal/Spatial X Analysis** – this extends Longitudinal X Analysis by going beyond simple before-after sequences and searching for full temporal (á la Allen [2] or Freksa [17]) or spatial (á la Egenhofer [14]) relationships between the patterns formed by the values or in the itemset or rule support, confidence or other rule qualifications. In extreme cases, this technique can suggest cause-and-effect.

2.3 External Datasets

As someone's health is often related to their environment, the accommodation of standard medical datasets (for example, the ICD10 and DRG standards for grouping diagnoses and procedures) is essential for sensible interpretation. As the semantics of the rules discovered are qualified by the semantics of the links between datasets, such linkages must be well understood. Note that the confidentialisation of data through the removal of keys can sometimes hinder such linking.

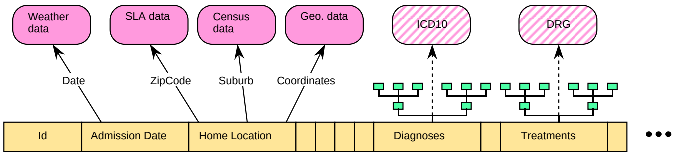

Figure 2: Linking in External and Hierarchy Files

However, as well as the standard medical datasets, we have found trusted external datasets (for example, statistical local area (SLA data), state and federal government data, meteorology data, and so on) to be particularly useful. Many contain hierarchies (see Section 2.5).

2.4 Rule Interpretation and Working with a Considerable Knowledge Base

The interpretation of the results of mining over medical datasets requires significant domain expertise. It is somewhat simpler for the developers of data mining routines to understand the results of routines run over, say, market basket or financial trading data. The need for strong cooperation between the IT researcher and the medical domain expert(s) is thus higher.

The results of data mining runs are never a set of 'rules' in the deductive sense and this is particularly true of medical data mining. Any results found should be interpreted solely as suggestions for future research. In many cases, the rules have been generally known (at least by the medical fraternity) and in other cases less widely, but still nevertheless known by specialists. Only in a few cases have our routines yielded genuinely new knowledge; we would argue, however, that the process is often worth the effort just for these few cases.

One of the major differences between medical data mining and mining over other fields is the vast (and assumed) knowledge existing in the field. In one study[3] the proportion of rules that were already readily known by the medical team members was almost total. In response to this, our itemset and rule visualisation algorithm $IsetNav$ was redeveloped to accommodate a knowledge base into which known rules could be placed leading to the (optional) suppression of these rules[4]. $IsetNav$ also accommodates hierarchies as discussed earlier.

[3]The results of this study (although not the technical mining details) are reported in [18].

[4]Note that in $IsetNav$ a rule is considered to consist not only of its antecedents and consequents but also of its support and confidence (ie. its full signature á la [41]). A significant change in, say, the confidence of a rule would thus be interpreted as interesting and the rule would be redisplayed.

2.5 Accommodating Multiple Hierarchies

While early association rule mining algorithms ran over attributes defined over simple (atomic) domains or domains with just one hierarchy, within medical domains it has become evident that many, if not most situations require the intelligent handling of multiple-hierarchy domains. For example, while a rule including a leaf node (say, a particular pathology test, disease or drug) may not have the required support, a higher level concept (a class of test, disease or the generic drug) might do and sensible handling is therefore required. Significantly, multiple-hierarchy domains, which require routines to traverse more than one hierarchy for each attribute, significantly increases the complexity of the algorithms.

Such extensions are particularly necessary when handling medical datasets (such those inherent in the ICDn and DRG standards) or datasets that may be incorporated to reveal medical associations (for example, postcode to SLA correlations, meteorological data, and so on). Current work in this general area includes [22; 23; 40] as well as work looking at multi-level association rules [5; 8; 19] although some of these do not readily lend themselves to multiple-hierarchy domains .

2.6 Data Availability and Accuracy

The datasets generated by medical practice and health research differ from those in a large number of the other areas within which data mining has been applied.

- **Encoding Problems.** While many medical datasets are generally of a high quality, and an increasing number of details are being recorded electronically, many are not encoded in a manner that makes them amenable for immediate use. Thus some data conversion and/or data cleaning can usually be expected [30; 43; 45]. This differs from either automatically collected datasets (such as market basket datasets) where data quality is near perfect and from survey/form collected data where the error rate can be relatively high. This often means that alternative mechanisms for outlier/error elimination may have to be accommodated[5].

- **Ethical and Legal Issues.** As discussed more fully in Section 2.7 the ethical requirements of handling even confidentialised medical data is significant. Relatively little work has been undertaken in this area although the notion of statistical compromise is well understood (see for example [28]) and a framework to alert users to sensitive rules has been proposed in [48].

- **Ownership.** Data ownership can easily stymie efforts at obtaining the data needed [10] or creating links between datasets. As discussed in Section 2.3, linking datasets is an important prerequisite to many epidemiological / demography / population health analyses and thus obtaining the necessary permissions can represent a large part of the preparation time.

- **Semantic Interoperability.** The semantic interoperability of data is an important issue. This issue is

[5]Note also that it is sometimes the outliers, or some of the smallest emergent or persistent clusters/rules that are of the greatest interest.

real and in some cases insuperable. In NSW, Australia, for example, each area health service has its own medical data standards which can cause interoperability difficulties. Work on data storage and interchange standards such as HL7, GEHR and others will, over time, assist in this.

While this is not an algorithmic or technical issue, much of the literature, as well our own experiences, have suggested that we ignore the complexities of data availability at our peril.

2.7 Ethical Safeguards and Sensitivity Alerts

The field of knowledge discovery, in common with many powerful technologies, lends itself both to abuse and to great benefit. However, unlike many technologies, the ability to compromise privacy, to stereotype, or simply to cause offense (potentially leading to litigation) can often be inadvertent. The complexities of the legal framework [39] together with the complexities of rule generation, particular medical rule generation, often means that the human post-processing of a data mining run can be a long and potentially complex process. As the focus is inevitably on the medical message being given, sociological issues in specific attribute values can be missed. For example, consider the following (simple and fictional) rule:

$$ZipCode(52409), Age(18-25), Gender(Male) \\ \rightarrow HIVStatus(Yes) \qquad \gamma(5\%)$$

This would be a worrying rule to discover for any segment of the population but consider the increased risk of offense if $ZipCode(52409)$ referred to, say, an indigenous community, a University campus or an area of high ethnic migration.

Both within and outside of the KDD community there is significant concern regarding the (ab)use of sensitive information [4; 6; 11; 12; 10; 32; 42]. Estivill-Castro *et al.*, for example, cite recent surveys about public opinion surrounding personal privacy which show concern about the use of private information such as medical data [15]. However, with some significant exceptions (see for example [1; 36]), to date, few practical solutions have been advanced.

A serious breach of privacy might result in a reactive legislative response which could hinder research such as this. We therefore suggest that, in additional to the normal ethical safeguards and existing government legislation, a potential research direction could be for the KDD community to develop a standard set of guidelines to assist researchers with these issues.

3. DISCUSSION AND FUTURE DIRECTIONS

From the previous section, as well as the specific medical data mining issues discussed, a number of future research directions in data mining seem to be highlighted. In summary these include those that:

- provide better support for the medical questions being asked (including changes to the investigative method), and those that
- enable the highly multi-dimensional nature of medical data to be accommodated (including handling external datasets, multiple-hierarchy domains, and the spatio-temporal nature of medical enquiry).

On top of these there are a variety of practical and ethico-legal considerations which could warrant some attention.

There are, of course, a number of areas of pertinent research we have not mentioned here, mainly because we have as yet had no experience of using them over medical data, although it is clear that their application could be useful. These include the mining of visual images (such as scans) and the visualisation of medical data and rules.

In summary, mining over medical, health or clinical data is arguably the most difficult domain for the KDD field. Indeed, for many techniques it can be safely stated that if the technique can be made to function usefully in the medical domain, it can be applied just about anywhere else. Moreover, while many of the issues discussed present distinct challenges for medical data mining, we believe that they might also be seen as indicators for future data mining research more generally.

4. REFERENCES

[1] R. Agrawal and R. Srikant. Privacy-preserving data mining. In W. Chen, J. Naughton, and P. A. Bernstein, editors, *ACM SIGMOD Conference on the Management of Data*, pages 439–450, Dallas, TX, 2000. ACM Press.

[2] J.F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.

[3] R.L. Blum. Discovery, confirmation and interpretation of causal relationships from a large time-oriented clinical database: The RX project. *Computers and Biomedical Research*, 15(2):164–187, 1982.

[4] C. Boyens, O. Gunther, and M. Teltzrow. Privacy conflicts in CRM services for online shops: A case study. In *[13]*, 2002.

[5] T. Calders, R.T. Ng, and J. Wijsen. Searching for dependencies at multiple abstraction levels. *ACM Transactions on Database Systems*, 27(3):229–260, 2002.

[6] A. Cavoukian. Data mining: Staking a claim on your privacy, 1998.

[7] P. Chapman, R. Kerber, J. Clinton, T. Khabaza, Thomas P. Reinartz, and Ruediger Wirth. The CRISP-DM process model. Discussion paper, CRISP-DM Consortium, 1999.

[8] D.W. Cheung, V.T. Ng, and B.W. Tam. Maintenance of discovered knowledge : A case in multi-level association rules. In *Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 307–310, Portland, Oregon, 1996. AAAI Press, Menlo Park, CA, USA.

[9] K.J. Cios, editor. *Medical Data Mining and Knowledge Discovery*, volume 60 of *Studies in Fuzziness and Soft Computing*. Physica-Verlag, New York, 2001.

[10] K.J. Cios and G.W. Moore. Medical data mining and knowledge discovery: Overview of key issues. In *[9]*, pages 1–20. 2001.

[11] R. Clarke. Privacy and dataveillance, and organisational strategy. In *Region 8 EDPAC'96 Information Systems Audit and Control Assoc. Conf*, Perth. Australia, 1997.

[12] R. Clarke. Person-location and person-tracking: Technologies, risks and policy implications. In *21st International Conference on Privacy and Personal Data Protection*, pages 131–150, 1999.

[13] C. Clifton and V. Estivill-Castro, editors. *Privacy, Security and Data Mining - Proc. IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining*, volume 14 of *Conferences in Research and Practice in Information Technology.* ACS, Maebashi City, Japan, 2002.

[14] M. Egenhofer and R. Franzosa. Point-set topological spatial relations. *International Journal of Geographical Information Systems*, 5(2):161–174, 1991.

[15] V. Estivill-Castro, L. Brankovic, and D.L. Dowe. Privacy in data mining. *Privacy - Law and Policy Reporter*, 6(3):33–35, 1999.

[16] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors. *Advances in Knowledge Discovery and Data Mining.* AAAI Press, Menlo Park, CA, USA, 1996.

[17] C. Freksa. Temporal reasoning based on semi-intervals. *Artificial Intelligence*, 54:199–227, 1992.

[18] A. Gialamas, J.J. Beilby, N.L. Pratt, R. Henning, J.E. Marley, and J.F. Roddick. Investigating tiredness in Australian general practice: Do pathology tests help in diagnosis? *Australian Family Physician*, 2003.

[19] J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. In U. Dayal, P.M.D. Gray, and S. Nishio, editors, *Twenty-first International Conference on Very Large Databases (VLDB '95)*, pages 420–431, Zurich, Switzerland, 1995. Morgan Kaufmann Publishers, Inc. San Francisco, USA.

[20] J. Han and M. Kamber. *Data Mining : Concepts and Techniques.* Academic Press, San Diego, 2001.

[21] A. Houston, H. Chen, S.M. Hubbard, B.R. Schatz, T.D. Ng, R.R. Sewell, and K.M. Tolle. 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, 1999.

[22] M. Kamber, J. Han, and J.Y. Chiang. Metarule-guided mining of multi-dimensional association rules using data cubes. In *Third International Conference on Knowledge Discovery and Data Mining*, pages 207–210. AAAI Press, Menlo Park, CA, USA, 1997.

[23] M. Kamber, J. Han, and J.Y. Chiang. Using data cubes for metarule-guided mining of multi-dimensional association rules. Technical Report CMPT-TR-97-10, School of Computing science, Simon Fraser University, May 1997.

[24] Y. Li, N. Wu, S. Jajodia, and X.S. Wang. Enhancing profiles for anomaly detection using time granularities. *Journal of Computer Security*, 10(1,2):137–157, 2002.

[25] J.M. Long, E.A. Irani, and J.R. Slagle. Automating the discovery of causal relationships in a medical records database. In G. Piatetsky-Shapiro and W.J. Frawley, editors, *Knowledge discovery in databases*, pages 465–476. AAAI Press/MIT Press, 1991.

[26] Ying Lu and Jiawei Han. Cancer classification using gene expression data. *Information Systems*, 28(4):243–268, 2003.

[27] M. McLeish, P. Yao, M. Garg, and T. Stirtzinger. Discovery of medical diagnostic information: An overview of methods and results. In G. Piatetsky-Shapiro and W.J. Frawley, editors, *Knowledge Discovery in Databases*, pages 477–490. AAAI Press/MIT Press, Cambridge, MA, 1991.

[28] M. Miller. A model of statistical database compromise incorporating supplementary knowledge. In B. Srinivasan and J. Zeleznikow, editors, *Second Australian Database-Information Systems Conference*, Sydney, Australia, 1991. World Scientific.

[29] C. Ordonez, C.A. Santana, and L. de Braal. Discovering interesting association rules in medical data. In D. Gunopulos and R. Rastogi, editors, *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery 2000*, pages 78–85, Dallas, Texas, 2000.

[30] J.C. Prather, D.F. Lobach, L.K. Goodwin, J.W. Hales, M.L. Hage, and W.E. Hammond. Medical data mining: Knowledge discovery in a clinical data warehouse. In *1997 Annual Conference of the American Medical Informatics Association*, Philadelphia, PA, 1997. Hanley and Belfus.

[31] G.M. Provan and M. Singh. Data mining and model simplicity: A case study in diagnosis. In E. Simoudis, J. Han, and U. Fayyad, editors, *Second International Conference on Knowledge Discovery and Data Mining (KDD96)*, Portland, Oregon, 1996. AAAI Press, Menlo Park, CA, USA.

[32] J. Rachels. Why privacy is important. *Philosophy and Public Affairs*, 4(4), 1975.

[33] C.P. Rainsford and J.F. Roddick. Adding temporal semantics to association rules. In J.M. Zytkow and J. Rauch, editors, *3rd European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'99)*, volume 1704 of *Lecture Notes in Artificial Intelligence*, pages 504–509, Prague, 1999. Springer.

[34] C.P. Rainsford and J.F. Roddick. Database issues in knowledge discovery and data mining. *Australian Journal of Information Systems*, 6(2):101–128, 1999.

[35] J.C.G. Ramirez, L.L. Peterson, D.M. Peterson, and G.K. Cormier. On the discovery of patterns in medical data. In *Fourteenth National Conference on Artificial Intelligence and Ninth Innovative Applications of Artificial Intelligence Conference, AAAI 97, IAAI 97*,

page 842, Providence, Rhode Island, 1997. AAAI Press / MIT Press.

[36] J.F. Roddick and P. Fule. A system for detecting privacy and ethical sensitivity in data mining results. Technical Report SIE-03-001, School of Informatics and Engineering, Flinders University, 2003.

[37] J.F. Roddick and B.G. Lees. Paradigms for spatial and spatio-temporal data mining. In *Geographic Data Mining and Knowledge Discovery*, Research Monographs in Geographic Information Systems. Taylor and Francis, 2001.

[38] J.F. Roddick and M. Spiliopoulou. A survey of temporal knowledge discovery paradigms and methods. *IEEE Transactions on Knowledge and Data Engineering*, 14(4):750–767, 2002.

[39] J.M. Saul. Legal policy and security issues in the handling of medical data. In *[9]*, pages 21–40. 2001.

[40] L. Shen and H. Shen. Mining flexible multiple-level association rules in all concept hierarchies (extended abstract). In G. Quirchmayr, E. Schweighofer, and T.J.M. Bench-Capon, editors, *Ninth International Conference on Database and Expert Systems Applications, DEXA'98*, volume 1460 of *Lecture Notes in Computer Science*, pages 786–795, Vienna, Austria, 1998. Springer-Verlag.

[41] M. Spiliopoulou and J.F. Roddick. Higher order mining: Modelling and mining the results of knowledge discovery. In N. Ebecken and C.A. Brebbia, editors, *Data Mining II - Proc. Second International Conference on Data Mining Methods and Databases*, pages 309–320. WIT Press, Cambridge, UK, 2000.

[42] B. Thuraisingham. Data mining, national security, privacy and civil liberties. *SIGKDD Explorations*, 4(2):1–5, 2003.

[43] V.S. Tigrani and G.H. John. Data mining and statistics in medicine: An application in prostate cancer detection. In *Proceedings of the Joint Statistical Meetings, Section on Physical and Engineering Sciences*, Dallas, TX, 1998. American Statistical Association.

[44] S. Tsumoto. Rule discovery in large time-series medical databases. In J. Zytkow and J. Rauch, editors, *Principles of Data Mining and Knowledge Discovery, Third European Conference, PKDD '99*, volume 1704 of *Lecture Notes in Computer Science*, pages 23–31, Prague, Czech Republic, 1999. Springer.

[45] S. Tsumoto. Discovery of clinical knowledge in databases extracted from hospital information systems. In *[9]*, pages 433–454. 2001.

[46] Anthony K. H. Tung, Hongjun Lu, Jiawei Han, and Ling Feng. Efficient mining of intertransaction association rules. *IEEE Transactions on Knowledge and Data Engineering*, 15(1):43–56, 2003.

[47] M.S. Viveros, J.P. Nearhos, and M.J. Rothman. Applying data mining techniques to a health insurance information system. In T.M. Vijayaramam, A. Buchmann, C. Mohan, and N.L. Sarda, editors, *Twenty-second International Conference on Very Large Data Bases (VLDB '96)*, pages 286–293, Mumbai, India, 1996. Morgan Kaufmann Publishers, Inc. San Francisco, USA.

[48] K. Wahlstrom and J.F. Roddick. On the impact of knowledge discovery and data mining. In John Weckert, editor, *Selected Papers from the Second Australian Institute of Computer Ethics Conference*, volume 1 of *Conferences in Research and Practice in Information Technology*. ACS, Canberra, 2001.