Intelligent data analysis in medicine

Nada Lavrač
J. Stefan Institute
Jamova 39
1001 Ljubljana
Slovenia

Elpida Keravnou
University of Cyprus
Kallipolenos 75
Nicosia
Cyprus

Blaž Zupan
J. Stefan Institute
Jamova 39
1001 Ljubljana
Slovenia

Abstract

Excessive amounts of knowledge and data stored in medical databases request the development of specialized tools for storing and accessing of data, data analysis, and effective use of stored knowledge and data. This paper first sketches the history of research that led to the development of current intelligent data analysis techniques, aimed at narrowing the increasing gap between data gathering and data comprehension. Next, we present our view on the relation of Intelligent data analysis to Knowledge discovery in databases and Data mining. Finally, we discuss the need for intelligent data analysis in medicine and present the aims of research in this area.

1 Introduction

"Now that we have gathered so much data, what do we do with it?" This is the opening statement of the editorial by Usama Fayyad and Ramasamy Uthurusamy in the Communications of the ACM, Special issue on Data Mining [12]. Recently, many statements of this kind appeared in journals, conference proceedings, and other materials that deal with data analysis, knowledge discovery, and machine learning. They all express a concern about how to "make sense" from the large volumes of data being generated and stored in almost all fields of human activity.

Especially in the last few years, the digital revolution provided relatively inexpensive and available means to collect and store the data. For example, in the domain of medicine, still in the mid-nineties one of the fathers of Artificial Intelligence in Medicine, Edward H. Shortliffe, partially blamed the underdeveloped hospital infrastructure for the failure to fulfill the initial promise of the field [38]. Recently, however, the situation is changing rapidly: modern hospitals are well equipped with monitoring and other data collection devices, and data is gathered and shared in inter- and intra-hospital information systems. In fact, medical informatics has become a must and an integral part of every successful medical institution [40].

The increase in data volume causes greater difficulties in extracting useful information for decision support. The traditional manual data analysis has become insufficient, and methods for efficient computer-based analysis indispensable. From this need, a new interdisciplinary field, that of Knowledge discovery in databases (KDD) was born [16]. KDD encompasses statistical, pattern recognition, machine learning, and visualization tools to support the analysis of data and discovery of principles that are encoded within the data.

The results of computer-based analysis have to be communicated to humans in an understandable way. In this respect, the analysis tools have to deliver transparent results and most often facilitate human intervention in the analysis process. A good example of such methods are symbolic machine learning algorithms that, as a result of data analysis, aim to derive a symbolic model (e.g., a decision tree or a set of rules) of preferably low complexity but high transparency and accuracy. Being in the core of KDD, the interest and research efforts in machine learning were largely increased.

This paper first sketches the history of research that led to the development of current intelligent data analysis techniques, aimed at narrowing the increasing gap between data gathering and data comprehension. Next, we present our view on the relation of Intelligent data analysis to Knowledge discovery in databases and Data mining. Finally, we discuss the need for intelligent data analysis in medicine and present the aims of research in this area.

2 Knowledge versus data: A historical sketch

In late seventies and early eighties, AI in medicine was mainly concerned with the development of medical expert systems aimed at supporting diagnostic decision making in specialized medical domains. Shortliffe's MYCIN [37], representing pioneering work in this area, was followed by numerous other efforts leading to specialized diagnostic and prognostic expert systems, e.g., HODGKINS [36], PIP [31; 41], CASNET [44], HEADMED [17], PUFF [22], CENTAUR [1], VM [11], ONCOCIN [39], ABEL [30], GALEN [42] MDX [8], and many others. The most general and elaborate systems were developed for supporting of diagnosis in internal medicine [32; 28; 27]: INTERNIST and its follower CADUCEUS, which, in addition to expert-defined rules as used in INTERNIST, included also a network of patophysiological states representing "deep" causal knowledge about the problem. The main problems addressed at this early stage of expert system research concerned knowledge acquisition [9; 10], knowledge representation, reasoning and explanation [43]. A typical early expert system schema is shown in Figure 1.

Rules were proposed from the early days of knowledge-based systems, and expert systems in particular, as a prime formalism for expressing knowledge in a symbolic way. Rules have the undisputed advantages of simplicity, uniformity, transparency, and ease of inference, that over the years have made them one of the most widely adopted approaches for representing real world knowledge. Rules elicited directly from domain experts are expressed at the right level of abstraction from the perspective of the expert, and are indeed comprehensible to the expert since they are formulations of his rules of thumb. However, human-defined rules risk capturing the biases of one expert, and although each rule individually may appear to form a coherent, modular chunk of knowledge, the analysis of rules as an integral whole can reveal inconsistencies, gaps, and the deficiencies due to their largely flat organization (i.e., the lack of a comprehensive, global, hierarchical organization of the rules).

It soon became clear that knowledge acquisition is the hardest part of the expert system development task. This was identified as the so-called "Feigenbaum bottleneck" [14; 15] in the construction of a knowledge base. The knowledge base is the heart of an expert system. For the effective use of expert system technology a knowledge base needs to be complete and consistent. The limitations of the first generation of expert systems [19; 23] coupled with the relatively high costs (in human and other terms) involved in acquiring knowledge directly from the experts, as well as the fact that databases of example cases started becoming readily available made the learning of rules from such data especially appealing, as a more efficient, less biased, and more cost-effective approach. On the one hand, this led to the developments in the area of machine learning (as described below), and on the other hand, to the investigations of the use of deep causal knowledge that could potentially overcome the difficulties encountered when using unstructured shallow-level sets of rules [20; 18]. An early approach in combining the use of deep knowledge and machine learning is presented in the KARDIO study [5].

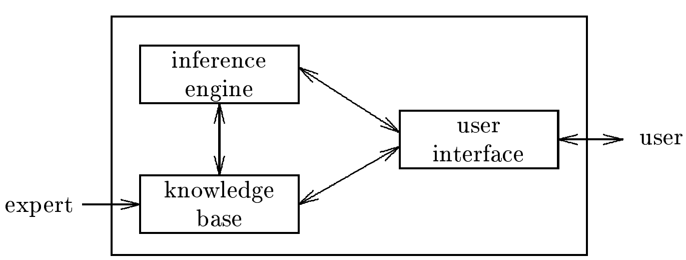

Figure 1: An expert system schema of early '80s.

In late eighties and early nineties it thus became apparent that knowledge acquired from experts alone is unsuitable for solving difficult problems and that, when developing decision support systems, the analysis of data gathered in the daily practice of experts and stored systematically in databases can play an important role in the support of expert decision making. This led to the development of early machine learning algorithms [25; 33] aimed at the automatic extraction of rules or decision trees from data. First machine learning systems aimed at dealing with real-world data which may be erroneous (noisy) and incomplete include CART [6], C4 [34], ASSISTANT [4; 7], and AQ [26], whereas C4.5 [35] is an efficient and the most popular machine learning system of the nineties.

Machine learning approaches do not advocate the bypassing of experts. Far from so. Experts are actively involved, but in a different and more constructive way than in the development of early expert systems. The example cases come from the experts and the resulting rules are validated by the experts for comprehensibility and other desired qualities. The learning approaches ensure that the derived rules are consistent, hierarchically organized (for example in terms of a decision tree), and, assuming that the collection of case examples used provides an adequate coverage of the particular domain, then the resulting set of rules will be of sufficient accuracy and adequate coverage (i.e., without significant gaps of knowledge). Furthermore, the expert provides important background knowledge for focusing and guiding the learning of rules. Irrespective of whether rules are learned or directly acquired from experts, their format should be simple, intuitive, and adequately expressive for the purposes of the particular application.

Most machine learning systems extract propositional rules from the data. Propositional rules are simple, categorical rules, with no variables, which nonetheless are

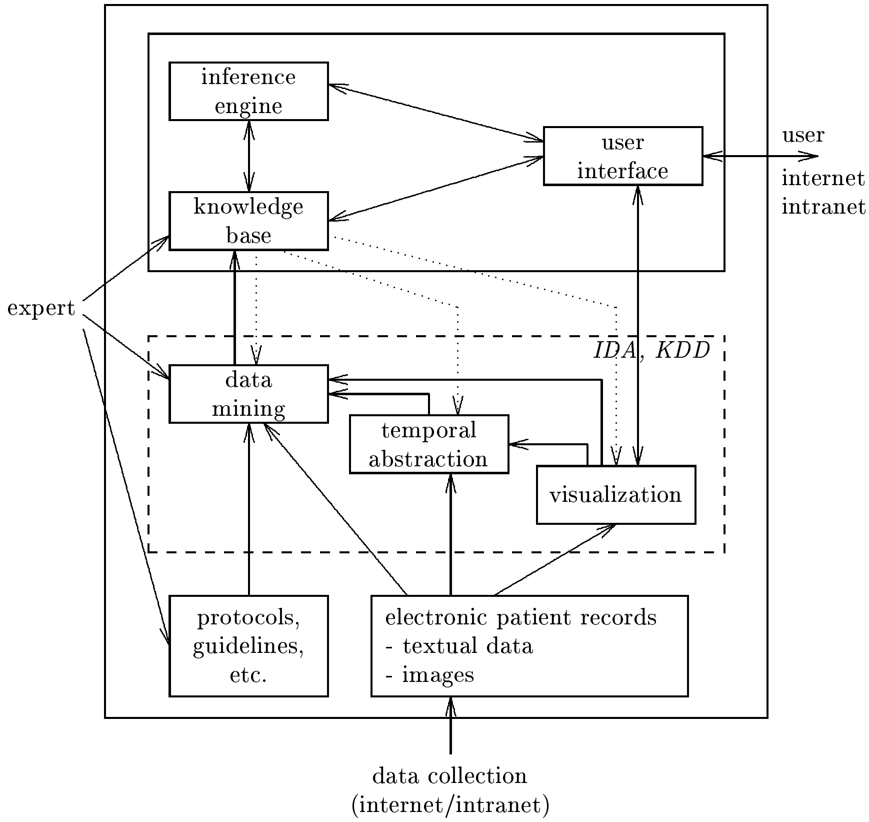

Figure 2: A decision support system schema of late '90s.

considered adequately expressive for many real applications. Predicate rules, on the other hand, are substantially more expressive and thus first-order learning techniques (recently known as Inductive Logic Programming (ILP) [29; 24]) have been developed for domains where the additional expressiveness is required. ILP techniques are also well suited for learning from relational databases. Moreover, they have the ability of dealing with elaborate expert knowledge that can easily be formulated in the form of first-order theories captured in the background knowledge to be used by the learner.

The nineties are characterized by the increasing gap between the excessive storage of uninterpreted data and the understanding of the data, and the need to overcome this gap by the effective use of data analysis techniques. The main emphasis of current research is thus on data analysis. This led to the challenging new research areas of Knowledge discovery in databases [16], Data mining, and Intelligent data analysis, in which machine learning techniques have a major role when the goal of data analysis is knowledge extraction. Current machine learning research is characterized by a shift of emphasis towards relational learning (ILP, [29; 24]) and more elaborate statistics applied in learning and evaluation methodologies. In data analysis, another trend is towards data abstraction and, in particular, towards temporal data abstraction [21] that can be viewed as a form of preprocessing for further data analysis. In the late nineties, data analysis has an increased role also due to the fact that data gathering is becoming distributed (e.g., telemedicine [2]), and that the analysis of such data is even more demanding. Figure 2 shows a possible schema of a decision support system of the nineties, where decision support needs to deal also with large volumes of data, as well as data gathering and analysis via the Internet and an intranet (see also the account by [3]).

3 Intelligent data analysis: Its relation to Knowledge discovery in databases and Data mining

In order to argue for the selection of the term "Intelligent data analysis" used in this paper, we need to give a definition of what we understand under this term, in relation to the more established and more frequently used terms: Knowledge discovery in databases and Data mining.

Knowledge discovery in databases (KDD) is frequently defined as a *process* [13] consisting of the following steps:

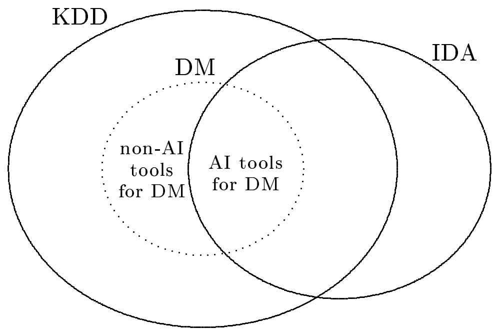

Figure 3: The relation of Intelligent data analysis (IDA), Knowledge discovery in databases (KDD), and Data mining (DM).

- understanding the domain,
- forming the dataset and cleaning the data,
- extracting of regularities hidden in the data thus formulating knowledge in the form of patterns, rules, etc.; this step in the overall KDD process is usually referred to as *Data mining* (DM),
- postprocessing of discovered knowledge, and
- exploitation of the results.

KDD is an interactive and iterative process in which many steps need to be repeatedly refined in order to provide for an appropriate solution to the data analysis problem. In this process, *data visualization* plays an important role. The *data mining* step in the KDD process deals with the extraction of knowledge from (typically large masses of) data, thus describing the data in terms of the interesting regularities discovered.

Intelligent data analysis (IDA) is largely related to KDD. It also refers to the interactive and iterative process of data analysis, with the distinguishing feature that the architectures, methodologies and techniques used in this process are those of Artificial Intelligence. Figure 3 shows the relation of IDA to Knowledge discovery in databases and Data mining.

There is a large intersection between KDD and IDA. The two fields have in common the topic of investigation, which is data analysis, and they share many common methods. However, there are some differences in the characteristics of applied methods and domains under investigation. As stated above, the main difference is that IDA uses AI methods and tools, while KDD employs both AI and non-AI methods (e.g., machine learning data mining techniques are in the intersection of the two fields, whereas classical statistical methods belong to KDD but not to IDA). Another aspect concerns the size of data: KDD is typically concerned with the extraction of knowledge from very large datasets, whereas in IDA the datasets are either large or relatively small. This also affects the type of data mining tools used: in KDD the data mining tools are executed mostly in batch mode (despite the fact that the entire KDD process is interactive), whereas in IDA the tools can either be batch or applied as interactive assistants.

4 IDA in medicine - Why is it needed ?

The gap between data generation and data comprehension is widening in all fields of human activity. In medicine overcoming this gap is particularly crucial since medical decision making needs to be supported by arguments based on basic medical knowledge as well as knowledge, regularities and trends extracted from data.

The significance of intelligent data analysis methods is manyfold:

- The first important aspect concerns the support of specific knowledge-based problem solving activities (diagnosis, prognosis, monitoring, etc.) by gleaning out the useful information from the raw, mostly numeric and often quite noisy, problem data; such data analysis methods are referred to as *data abstraction methods*. Here, the prime type of intelligent data analysis is the derivation of temporal trends, and overall the *temporal* dimension is of a crucial concern. Since the primary goal of these methods is (on-line) decision support, their quality assessment is performance-based: for instance, is a method accurate enough to be used for diagnosis and prediction, does it predict well a trend or a value to be expected at the next point of time? In this context, *visualization of data* is extremely important for supporting decision making and even invaluable for successfully performing a problem solving task.
- The second aspect concerns the discovery of new medical knowledge that can be extracted through data mining of representative collections of example cases, described by symbolic or numeric descriptors. The available datasets are mostly incomplete (missing data) and noisy (erroneous). The methods for extracting meaningful and understandable symbolic knowledge will be referred to as *data mining methods*. The quality assessment of these methods is based both on the performance (classification and prediction accuracy, misclassification cost, sensitivity, specificity, etc.), as well as the understandability and significance of the discovered knowledge.

5 Towards a classification of IDA methods

Based on the main aspects of the use of IDA methods in medicine discussed in Section 4, we propose the following classification of IDA methods:

- *Data abstraction methods* are intended to support specific knowledge-based problem solving activities (data interpretation, diagnosis, prognosis, monitoring, etc.) by gleaning out the useful abstractions from the raw, mostly numeric problem data. *Temporal data abstraction methods* represent an important subgroup where the processed data are temporal. The derivation of abstractions is often done in a context sensitive and/or distributed manner and it applies to discrete and continuous supplies of data. Useful types of temporal abstractions are trends, periodic happenings, and other forms of temporal patterns. Temporal abstractions can also be discovered by visualization. The abstraction can be performed over a single case (e.g., a single patient) or over a collection of cases.

- *Data mining methods* are intended to extract knowledge preferably in a meaningful and understandable symbolic form. Most frequently applied methods in this context are supervised symbolic learning methods. For example, effective tools for inductive learning exist that can be used to generate understandable diagnostic and prognostic rules. Symbolic clustering, discovery of concept hierarchies, qualitative model discovery, and learning of probabilistic causal networks fit in this framework as well. Sub-symbolic learning and case-based reasoning methods can also be classified in the data mining category. Other frequently applied sub-symbolic methods are nearest neighbor, Bayesian classifier and non-symbolic clustering.

6 Goals of IDA research in medicine

In this section we summarize the targets for Intelligent data analysis in medicine. Any research in medicine aims to directly or indirectly enhance the provision of health care. IDA research in this fields is no exception. The goals of IDA in medicine are:

- the intelligent interpretation of patient data in a context-sensitive manner and the presentation of such interpretations in a visual or symbolic form; the temporal dimension in the representation and intelligent interpretation of patient data is of primary importance,

- the extraction (discovery) of medical knowledge for diagnostic, screening, prognostic, monitoring, therapy support or overall patient management tasks.

IDA research in medicine is driven by a very pragmatic aim, the enhancement of health care, and as such the benchmark tests for such methods and techniques can be no less than real world problems. Viable IDA proposals for medicine must be accompanied by detailed requirements that delineate the spectrum of real applications addressed by such proposals, and the in-depth evaluation of resulting systems constitutes a critical aspect.

Another consideration is the role of IDA systems in a clinical setting. Their role is clearly that of an intelligent assistant that tries to bridge the gap between data gathering and data comprehension, in order to enable the physician to perform his task more efficiently and effectively. If the physician has at his disposal the right information at the right time then doubtless he will be in a better position to reach correct decisions or instigate correct actions in a timely fashion. The information revolution has made the collection and storage, on electronic media, of large volumes of data from diverse sources a possibility. These data can be on a single case (e.g., one patient) or on multiple cases. Raw data as such are of little value since their sheer volume and/or the very specific level at which they are expressed renders their utilization (operationalization) in the context of problem solving an impossibility. However such data can be converted to a mine of information wealth if the real gems of information are gleaned out by computationally intelligent means. The useful, operational information/knowledge, which is expressed at the right level of abstraction, is then readily available to support the decision making of the physician in managing a patient.

Important issues that arise from the rapidly emerging globality of data and information are:

- the provision of standards in terminology, vocabularies and formats to support multi-linguality and sharing of data,

- standards for the abstraction and visualization of data,

- standards for interfaces between different sources of data,

- seamless integration of heterogeneous data; images and signals are important types of data,

- standards for electronic patient records, and

- reusability at all levels, i.e., data, knowledge and tools.

Clinical data constitute an invaluable resource for the human race, the proper utilization of which impinges directly on the essential aim of health care which is "correct patient management". Investing in the development of appropriate IDA methods, techniques and tools for the analysis of clinical data is thoroughly justified and this research ought to form a main thrust of activity by the relevant research communities.

7 Conclusion

Intelligent data analysis is a recently emerging research area that was born from the need of extracting useful information from increased volumes of data gathered in databases. In medicine, overcoming the gap between data gathering and interpretation is particularly crucial.

The goal of this paper is to relate Intelligent data analysis research with respect to the past and research trends, in particular to Knowledge Discovery in Databases and Data mining. The paper proposes a classification of Intelligent data analysis methods into data abstraction and data mining methods, and discusses the needs for Intelligent data analysis in medicine.

Acknowledgments

Nada Lavrač and Blaž Zupan are supported by the Ministry of Science and Technology of Slovenia. The support for Elpida Keravnou is provided by a research grant from the University of Cyprus.

References

[1] Aikins, J. S. (1979). Prototypes and production rules: An approach to knowledge representation for hypothesis formation. In *Proc. Sixth International Joint Conference on Artificial Intelligence*, pages 1–3.

[2] Barahona, P., and Christensen, J.P., editors (1994). *Knowledge and Decisions in Health Telematics*, IOS Press.

[3] van Bemmel, J. H. (1996). Medical informatics, art or science? *Meth. Inform. Med.*, 35:157–172.

[4] Bratko, I. and Kononenko, I. (1987). Learning diagnostic rules from incoplete and noisy data. In Phelps, B., editor, *AI Methods in Statistics.* Gower Technical Press, London.

[5] Bratko, I., Mozetič, I., and Lavrač, N. (1989). *KARDIO: A Study in Deep and Qualitative Knowledge for Expert Systems.* MIT Press, Cambridge, MA.

[6] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth, Belmont.

[7] Cestnik, B., Kononenko, I., and Bratko, I. (1987). Assistant 86: A knowledge-elicitation tool for sophisticated users. In Bratko, I., and Lavrač, N., editors, *Progress in Machine Learning*, Sigma Press, pages 31–45.

[8] Chandrasekaran, B., and Mittal, S. (1983). Conceptual representation of medical knowledge for diagnosis by computer: MDX and related systems. *Advances in Computers*, 22:217–293.

[9] Davis, R. (1978). Knowledge acquisition in rule-based systems: knowledge representation as a basis for system construction and maintainance. In Shine, M. E. and Coombs, M. J., editors, *Designing for Human-Computer Communication*, pages 87–137. Academic Press.

[10] Davis, R. (1983). TEIRESIAS: experiments in communicating with a knowledge-based systems. In Waterman, D.A., and Hayes-Roth, F., editors, *Pattern Directed Inference Systems*, pages 99–134. Academic Press.

[11] Fagan, L. M., Shortliffe, E. H., and Buchanan, B. G. (1980). Computer-based medical decision making: From MYCIN to VM. *Automedica.*

[12] Fayyad, U. M., and Uthurusamy, R. (1996). Data minning and knowledge discovery in databases (editorial). *Communications of the ACM*, 39(11):24–26.

[13] Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996). The KDD process for extracting useful knwoledge from volumes of data. *Communications of the ACM*, 39(11):27–41.

[14] Feigenbaum, E. A. (1977). The art of artificial intelligence 1: Themes and case studies of knowledge engineering. Technical report, Pub. no. STAN-SC-77-621, Stanford University, Department of Computer Science.

[15] Feigenbaum, E. A. and McCurduck, P. (1984). *The fifth generation.* Pan Books, London.

[16] Frawley, W., Piatetsky-Shapiro, G., and Matheus, C. (1991). Knowledge discovery in databases: An overview. In Piatetsky-Shapiro, G. and Frawley, W., editors, *Knowledge discovery in databases.* The AAAI Press, Menlo Park, CA.

[17] Haiser, J. F., Brooks, R. E., and Ballard, J. P. (1978). Progress report: A computerized psychopharmacology advisor. In *Proc. 11th Collegium Internationale Neuro-Psychopharmatologicum*, Vienna.

[18] Keravnou, E. T., editor (1992). *Deep models for medical knowledge engineering.* Elsevier.

[19] Keravnou, E. T., and Washbrro, J. (1989). What is a deep expert system: an analysis of the architectural requirements of second generation expert systems. *Knowledge Engineering Review*, 4(3):205–233.

[20] Keravnou, E. T., and Washbrook, J. (1989). Deep and shallow models in medical expert systems. *Artificial Intelligence in Medicine*, 1(1):11–28.

[21] Keravnou, E. T., editor (1996). Special issue on temporal reasoning in medicine. *Artificial Intelligence in Medicine*, 8(3):187–326.

[22] Kunz, J. C., et al. (1978). A physiological rule-based system for interpreting pulmonary function test results. Technical report, Stanford HPP Memo HPP-78-19.

[23] Lavrač, N. Mozetič, I. (1992) Second generation knowledge acqusition methods and their application to medicine. In Keravnou, E. T., editor, *Deep models for medical knowledge engineering*, pages 177–199, Elsevier.

[24] Lavrač, N. and Džeroski, S. (1994). *Inductive logic programming: Techniques and applications.* Ellis Horwood.

[25] Michalski, R. (1983). A theory and methodology of inductive learning. In Michalski, R., Carbonell, J., and Mitchell, T., editors, *Machine Learning: An Artificial Intelligence Approach*, volume I, pages 83–134, Palo Alto, CA. Tioga.

[26] Michalski, R., Mozetič, I., Hong, J., and Lavrač, N. (1986). The multi-purpose incremental learning system AQ15 and its testing application on three medical domains. In *Proc. Fifth National Conference on Artificial Intelligence*, pages 1041–1045, San Mateo, CA. Morgan Kaufmann.

[27] Miller, R. A. (1984). Internist-I/CADUCEUS: Problems facing expert consultant programs. *Meth. Inform. Med.*, 23:9–14.

[28] Miller, R. A., Pople, H. E., and Myers, J. D. (1982). Internist-I, An experimental computer-based diagnostic consultant for general internal medicine. *The New England Journal of Medicine*, 307(8):468–476.

[29] Muggleton, S. (1991). Inductive logic programming. *New Generation Computing*, 8(4):295–318.

[30] Patil, R. S., Szolovits, P., and Schwartz, W. B. (1982). Modelling knowledge of the patient in acid-base and electrolyte disorders. In Szolovits, P., editor, *Artificial Intelligence in Medicine*, AAAS Selected Symposium Series, pages 345–348. West View Press.

[31] Pauker, S. G., Gorry, G. A., Kassirer, J. P., and Schwartz, W. B. (1976). Towards the simulation of clinical cognition: Taking a present illness by computer. *The American Journal of Medicine*, 60:981–995.

[32] Pople, H. E. (1982). Heuristic methods for imposing structure on ill structured problems: The structuring of medical diagnosis. In Szolovits, P., editor, *Artificial Intelligence in Medicine*, AAAS Selected Symposium Series, pages 119–185. West View Press.

[33] Quinlan, J. R. (1983). Learning efficient classification procedures and their application to chess end-games. In Michalski, R. S., Carbonell, J. G., and Mitchell, T. M., editors, *Machine Learning: An artificial intelligence approach.* Tioga Publishing Company, Paolo Alto.

[34] Quinlan, J. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.

[35] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers.

[36] Safrans, C., Desforges, J., and Tsichlis, P. (1976). Diagnostic planning and cancer management. Rep. No. TR-169. Technical report, Laboratory for Computer Science, M. I. T., Massachusetts.

[37] Shortliffe, E. H. (1976). *Computer-Based Medical Consultations: MYCIN.* Elsevier.

[38] Shortliffe, E. H. (1993). The adolescence of AI in medicine: Will field come of age in the '90s? *Artificial Intelligence in Medicine*, 5(2):93–106.

[39] Shortliffe, E. H., Scott, C. A., and Bischoff, M. B. (1981). ONCOCIN: An expert system for oncology protocol management. In *Proc. Seventh International Joint Conference on Artificial Intelligence*, pages 876–881.

[40] Spackman, K., Elert, J. D., and Beck, J. R. (1993). The CIO and the medical informaticist: alliance for progress. In *Proc. Annual Symposium on Computer Applications in Medical Care*, pages 525–528.

[41] Szolovits, P. and Pauker, S. G. (1978). Categorical and probabilistic reasoning in medical diagnosis. *Artificial Intelligence*, 11.

[42] Thompson, W. B., Johnson, P. E., and Moen, J. B. (1983). Recognition-based diagnostic reasoning. In *Proc. Eight International Joint Conference on Artificial Intelligence*, pages 236–238.

[43] Wallis, J. W., and Shortliffe, E. H. (1982). Explanatory power of medical expert systems: studies in the representation of causal relationships for clinical consultatins. *Meth. Inform. Med.*, 21:127–136.

[44] Weiss, S. M., Kulikowski, C. A., Amarel, S., and Safir, A. (1978). A model-based method for computer-aided medical decision making. *Artificial Intelligence*, 11:145–172.