Abstract

Use of computer based decision tools to aid clinical decision making, has been a primary goal of research in biomedical informatics. Research in the last five decades has led to the development of Medical Decision Support (MDS) applications using a variety of modeling techniques, for a diverse range of medical decision problems. This paper surveys literature on modeling techniques for diagnostic decision support, with a focus on decision accuracy. Trends and shortcomings of research in this area are discussed and future directions are provided.

The authors suggest that – i) Improvement in the accuracy of MDS application may be possible by modeling of vague and temporal data, research on inference algorithms, integration of patient information from diverse sources and improvement in gene profiling algorithms; ii) MDS research would be facilitated by public release of de-identified medical datasets, and development of opensource data-mining tool kits; iii) Comparative evaluations of different modeling techniques are required to understand characteristics of the techniques, which can guide developers in choice of technique for a particular medical decision problem; iv) Evaluations of MDS applications in clinical setting are necessary to foster physicians’ utilization of these decision aids.

1 Introduction

MDS applications aim to assist physicians in clinical decision making [1]. Physicians are prone to making decision errors, because of high complexity of medical problems and due to cognitive limitations. Decision making in medicine is complex because, a vast amount of knowledge is required even to solve seemingly simple problems [2]. A physician is required to remember and apply knowledge of a large array of entities like disease presentations, diagnostic parameters, drug combinations and guidelines [3]. However the physician’s cognitive abilities are restricted due to factors like multi-tasking, limited reasoning and memory capacity [4, 5]. Consequently, it is impossible for an unaided physician to make right decisions every-time [6]. Ironically, the increasing rate of information generation by medical advances has aggravated the physician’s task [7, 8]. Humans have a “limited channel capacity” [4], due to which they are unable to process large amount of data. These limitations result in an estimated 30% morbidity and mortality [9, 10], which is avoidable.

The potential of computer based tools for medical decision making was realized half a century ago [11], and several algorithms have been developed to construct MDS applications for a variety of medical specialities. Figure 1 shows the logical components of the diagnostic process, as outlined by Ledley and Lusted [12]. Accordingly, MDS applications typically have two components – medical knowledge-base and an inference engine. Medical knowledge is composed of symptom-disease, inter-symptom and inter-disease relationships. The inference engine essentially interprets patient information using medical knowledge to compute decisions.
Despite half a century of research [13, 14], MDS applications have not received wide acceptance and utilization in health-care, due to the following reasons:

- **Decision accuracy:** High level of decision accuracy is desirable in MDS and very few applications have matched the diagnostic performance of human experts [15, 16, 17].

- **Integration with workflow:** The tools do not provide all the different types of information that the clinicians need [7, 18]. For instance, a diagnostic tool often does not provide the plan of treatment and the physician has to seek this information from other tools, which interrupts his workflow.

- **Usability:** The interfaces of many of the early systems were not user friendly and training was required for their use. Usability issues limited their use [19, 20].

- **Natural language processing:** Most patient information is unstructured and in a free textual form. Such data is first parsed using Natural Language Processing (NLP) algorithms to identify the parts of speech like noun, adjective, verb, etc. Ontologies are then used to perform semantic interpretation of the mapped words. NLP methods are still not adequate for clinical applications [21].

- **Explanation:** If scientific evidence is provided to explain the basis of decision-suggestions, physicians are more likely to use the system [22, 23]. Many systems have lacked such explanation modules.

Decision accuracy is a primary factor determining the adoption and utilization of MDS applications in routine clinical practice. This paper presents a survey of modeling techniques for diagnostic decision support, with a focus on decision accuracy. Though MDS modeling techniques have been applied for a range of medical decisions including diagnosis, prognosis and therapy planning, we have restricted the scope of the survey to the diagnostic decision problem. Diagnosis is the most researched decision, as diagnostic decisions are comparatively easier to validate. Moreover, most medical decision problems are classification problems and results for diagnostic modeling techniques can be usually extrapolated to other problems.

The aim of this survey is to provide an overview of modeling techniques for diagnostic decision support, along with a listing of their applications, and to provide a context and future directions to readers who are new to the field. The survey is organized by grouping literature on the basis of modeling techniques. Early MDS systems were largely bayesian, and were followed by fuzzy set theory (FST), rule-based and heuristic system. Recently Bayesian Network (BN), Artificial Neural Network (ANN) and Support Vector Machine (SVM) models have been developed (see figure 8). The next section describes the approach for the literature survey, followed by an overview of literature of different modeling techniques. The discussion section indicates trends, limitations and future directions.

2 Methods

We queried PubMed for *Computer assisted Diagnosis* (see figure 2) and largely excluded algorithms for image, waveform and -omic data analysis, because each of these areas is a specialized field in itself and applications in these domains generally cannot be ported to other problem areas. PubMed search yielded a total of 2880
papers. After excluding literature from the specialized areas mentioned above, we analyzed this set to delineate papers that reported either a new decision model or evaluation of existing models. These were then segregated into different groups based on the modeling technique used in the studies. Some reports were left out due to paper length restrictions, yielding the final set of 220 reports cited in this survey.

A limitation of our literature survey is that, our search largely relied on PubMed. Many conference proceedings and journals are not indexed in PubMed and hence, our survey is limited to a subset of published literature.


Figure 2: PubMed Query

3 Results

This section is organized by grouping literature on the basis of modeling techniques. Early MDS systems were largely bayesian, and were followed by FST, rule-based and heuristic system. Recently BN, ANN and SVM models have been developed (see figure 8).

3.1 Naive Bayes (NB)

Ledley and Lusted’s [12] paper outlining the use of Bayes formula to estimate probability of a diagnosis stimulated great interest. This approach is based on the assumption that the findings/symptoms for a particular disease are not interdependent. For example, the probability that a patient with symptoms of cough and fever has a diagnosis of pneumonia is computed as:

$$P(\text{pneumonia}) \times \frac{P(\text{cough}|\text{pneumonia}) \times P(\text{fever}|\text{pneumonia})}{P(\text{cough} \cap \text{fever})}$$  (1)

where $P(\text{cough}|\text{pneumonia})$ indicates probability of cough given the diagnosis of pneumonia. The assumption made for simplification, here is that there is no dependence between fever and cough.

One of the earliest studies with this approach was by Homer Warner at University of Utah. Warner made a probabilistic model to diagnose patients with one of 35 congenital heart diseases [24]. The model was derived on how frequently 50 different findings occurred in each disease and how common the diseases were in the population of patients referred to his laboratory. Warner’s research led to the development of HELP [25], which was the first hospital information system with decision support modules. Its decision support functions have been expanded over the years to provide alerts/reminders, data interpretation, patient diagnosis, patient management suggestions and clinical protocols. An offshoot of HELP was a sequential Bayes model of Iliad [26], developed for diagnosis in the domain of internal medicine.

In 1970s, De Dombal developed a system for diagnosis of pain in abdomen at Leeds University, UK. This was one of the first systems to be utilized at clinical sites and is perhaps the most evaluated system, with studies in UK and over 64 European institutions, together consisting of 37,000 cases [27]. These studies confirmed that MDS systems could outperform unaided clinicians, but aided clinicians matched and at times exceeded the diagnostic accuracy of the systems.

In 1987, DXPLAIN for internal medicine [28], was developed by Barnett et al. at Massachusetts General Hospital. It is currently available as a web-based system and includes 2200 diseases and 5000 symptoms in its knowledge-base. DXPLAIN produces a ranked list of diagnoses for clinical manifestations of a query patient and also provides justification for why each of these diagnoses might be considered and suggests what further clinical information would be useful to collect to further resolve the problem.

Modifications to Bayes rule to overcome the independence assumption, have been applied for diagnosis of acute abdominal pain [29], sport injuries [30], disease outbreak [31], lung disease [32] and human malformation patterns [33, 34]. MEDAS system at Chicago medical school used a modified Bayes approach [35].
Other early applications of the bayesian approach include diagnosis for hematologic disorders (HEME) [36], and appendicitis [37]. Later applications have been developed for abdominal pain [38, 39, 40, 41, 42], odontogenic lesions [43], oncology [44, 45, 46], liver disorders [47, 48, 49, 50, 51, 52], pancreatitis [53], lung disease [54, 55], dentistry [56, 57], gynecology [58, 59], neurology [60], rheumatology [61, 62], dermatopathology [63, 64], ophthalmoiology [65], hematopathology [66, 67], hypertension [68], heart disease [69], adverse drug reactions [70] and bowel pathology [71, 72, 73].

Many naive bayes (NB) based applications achieved a performance which was comparable to human physicians and some were successfully deployed in hospitals [27, 25, 28]. However, the assumptions of independence of symptoms and mutual exclusivity of diseases were regarded as over-simplistic, which led researchers to explore other approaches.

### 3.2 Decision Analysis

Decision analysis refers to “choice under uncertainty, that is consistent with a set of judgments and preferences for consequences”. The preferences for consequences are in terms of utility values, and judgments about uncertainties are modeled as probabilities. In clinical medicine, the analysis consists of determining the best action (diagnostic or therapeutic) for any given patient [74]. The physician considers the risks and costs of making the decisions. For instance, patients complaining of chest pain are routinely screened with Electrocardiogram (ECG) examination for myocardial infarction (heart-attack). Of the many causes for chest pain, myocardial infarction is one of the less likely ones, but entails high risk of fatality. Hence, the cost of a missed diagnosis of myocardial infarction is much higher than other causes.

In 1970, Ginsberg [75] and Gorry [76] implemented decision theory model for the diagnosis and treatment of pleural effusion syndrome and acute renal failure, respectively. Later PIP (Present Illness Program) was a system built by MIT and Tufts-New England Medical Center, for diagnosing renal disease. It used both categorical as well as probabilistic criteria for computing diagnosis [77]. After considerable groundwork to evaluate several approaches, in 1992 Heckerman developed a normative (Pathfinder) system [78] to assist surgical pathologists with the diagnosis of lymph-node diseases. The system used probability and decision theory to acquire, represent, manipulate, and explain uncertain medical knowledge. Research for Pathfinder led to an understanding of the reasons for the failure of some early Artificial Intelligence (AI)-based approaches.

Although physicians use a decision analysis based approach for making clinical decisions, utility considerations have been rarely used in MDS applications. A correction in MDS applications’ output by using decision analysis, may facilitate clinical utility of the applications.

### 3.3 Rule-based

Rule-based systems emphasize on use of decision rules for inference, as opposed to scoring functions. One of the first and widely known rule-based systems was MYCIN, developed by Shortliffe [79] at Stanford in 1977. Clinical knowledge in this application was represented as a set of if-then rules with certainty factors attached to consequent diagnoses (see figure 3). A backward chaining reasoning strategy was used for inference.

**IF** the infection is primary bacteremia
**AND** the site of the culture is one of the sterile sites
**AND** the suspected portal of entry is the gastrointestinal tract
**THEN** there is suggestive evidence (0.7) that infection is bacterial.

*Figure 3: Mycin rule*

Another rule-based system (EXPERT) [80], was developed by Weiss. An EXPERT model consisted of hypotheses (structured as causal and taxonomic networks), findings or observations, and decision rules for logically relating these components. The system was used for rheumatology, ophthalmology [81], and endocrinology. In 1983, Peter Politakis [82] developed SEEK system for giving interactive advice about rule refinement during the development of an expert system. It was used for dermatology [83].
In 80s and 90s, rule-based systems were developed for esophageal cancer [84], psychiatry (Psyxpert: using certainty and importance measures) [85, 86], ascites (using ID3 for rule induction)[87], female urinary incontinence [88], pregnancy disorders [89], kidney disease [90], metabolic disease [91], nosocomial infections [92], hepato-biliary disorders (HEPAR) [93], neurology [94, 95, 96], appendicitis [97], lung disease [98], thyroid disorder [99], pathology [100], uterine disease [101] and heart disease [102].

A disadvantage of rule-based applications is that they become increasingly unmanageable when the number of rules increase [102]. Methods for automated discovery and optimization of the rules have revived interest in the technique [103, 104]. Rule-based systems are inherently amenable to derivation of decision explanations.

### 3.4 Heuristic

Heuristic algorithms are those for which there is no formal proof of correctness. These include most of the AI programs developed in the 1980s and also include rule-based systems which used ad-hoc scoring schemes for inference. CASENET developed by Kulilowski was probably the first heuristic system. It used a descriptive model of the disease process for logical interpretation of clinical findings for glaucoma [105]. The causal model representing pathophysiological mechanisms had the form of a semantic net with weighted links. It was used to develop consultation systems for neuro-opthalmology, infectious diseases of the eye, rheumatology [106, 107] and pathology, and was extended into the natural language system called CHRONOS [108]. Around the same time, Graham developed a scoring index for gangrenous and perforating appendicitis [37].

Internist [17] was a wide-domain heuristic system developed for general medicine, at university of Pittsburgh. It had a knowledge-base of 600 diseases and 4500 findings, developed by a team of physicians and medical students, who reviewed literature to estimate for each finding, a frequency weight (FW) and evoking strength (ES) in grade range 1 – 5. FW was the frequency with which the finding was found in a disease and ES reflected the belief that a patient had a disease given that the finding was present. In addition, each finding had an import number (1-5), which reflected the importance of the finding. Internist used a scheme similar to the hypothetico-deductive approach for inference. The medical knowledge-base of INTERNIST was used to construct CADUCEUS [109] and a commercial system named Quick Medical Reference (QMR) [110] . CADUCEUS was designed to infer on cases with simultaneous diseases, and on flawed and scarce data. Inference engine of CADUCEUS was similar to MYCIN, but in addition incorporated abductive reasoning.

Other heuristic systems are for diagnosis in geriatric psychiatry (Methuselah) [111], rheumatology [112], cardiology [113] and malformation syndromes [114]. Despite the lack of formal proof of correctness, heuristic methods have found wide applications. Lack of mathematical rigour lead researchers to explore other modeling techniques.

### 3.5 Fuzzy Set theory

Fuzzy sets extend the concept of conventional sets, by allowing its elements to have degrees of membership. They are useful for representation of vague concepts expressed in natural language and uncertainties in measurement. The earliest suggestion of the applicability of FST to medicine is ascribed to Zadeh [115]. Early applications were largely based on fuzzy relations. However in the last decade, fuzzy rule-based systems have been popular and several other fuzzy set theory based approaches have been developed, as described below.

#### 3.5.1 Fuzzy Relations (FR)

Sanchez modeled medical knowledge as fuzzy relations (figure 4) between symptoms and diseases [116] and used compositional rule for inference [117]. Sanchez’s model was extended to develop CADIAG systems. CADIAG knowledge-base was constructed in the form of fuzzy relations acquired from physicians, medical literature and semi-automatically from patient records. CADIAG captured uncertainty in information about a patient’s symptoms and medical knowledge, as binary fuzzy relations, and it used a heuristic algorithm for diagnostic inference. These systems were evaluated to have high accuracy for rheumatic [118] and pancreatic [119] diseases. Recently Ciabattoni et al. [120] have defined a formal logical calculus to perform a consistency check of the CADIAG knowledge-base. Applications of fuzzy relations include diagnosis of cardiopathy [121] and arteritis.
Recently, Wagholikar et al. evaluated a fuzzy relation (FR) algorithm on a variety of medical datasets, and concluded that their algorithm is particularly useful for noisy and incomplete datasets.

`Probability of finding fever in patient having malaria is 0.8' translates to
'Fuzzy occurance relation of fever with malaria is 0.8'

`Probability of diagnosing malaria in patient who has fever is 0.2' translates to
'Fuzzy confirmatory relation of fever with malaria is 0.2'

3.5.2 Fuzzy rules

In these systems, the knowledge-base is generally derived as rule by inducing decision tree from data (figure 5). The crisp rules are then fuzzified by computing fuzzy sets. Such systems have been applied for aphasia [124], sleep apnea [103], coronary disease [125], post-operative infections [126], hepatic disorder [127], and malaria [128].

If temperature is ‘high’ and lymphocyte count is ‘high’, then confidence for viral infection is 0.95

3.5.3 Neuro-Fuzzy

Neuro-fuzzy systems combine FST with the theory of ANN. Kuncheva [129] first proposed the model of a fuzzy neuron for aviation medicine. Barreto [130] interpreted the connection values and excitation state of the neurons. Neuro-fuzzy approach has been applied for diagnosing erythematous-squamous diseases [131], heart disease [132] and hemorrhage [133].

3.5.4 Other FST based

Artificial Intelligence Research Institute, Spain [134] developed Milord system based on approximate reasoning for rheumatology (Renoir) and pneumonia (Pneumon-IA). Other FST applications include diagnosis of postmenopausal osteoporosis [135], myocardial ischemia [136], hematological disorders [137, 138], sore throat [139], fuzzy-genetic approach [140], fuzzy classification method [141, 142], fuzzy pattern recognition [143], fuzzy cognitive maps for pathology [144, 145], fuzzy similarity classifier [146] for erythematous-squamous diseases, intuitionistic fuzzy sets [147] and fuzzy influence diagrams [148].

FST offers a wide variety of formalisms for modeling medical decision making and is a fertile ground for MDS research [149]. However, the use of fuzzy models generally entails high computational costs.

3.6 Artificial Neural Network

Artificial Neural Network (ANN) is a mathematical model that simulates biological neural networks. It consists of an interconnected group of artificial neurons (nerve cells) that process information (figure 6). The strength/weight of the neuronal interconnections, topology of the network and choice of excitation function of the neurons are important factors governing the behavior and utility of the model. The applications of ANN to medicine appeared in early 1990s [150] and rapidly gained popularity [151]. ANNs have been applied for cardiology [152, 153, 154, 155], orthopedics [156], psychiatry [157], low back pain [158], cancer [159, 160, 161, 162], dyspepsia [163], atrophic gastritis [164], infection [165, 166, 167], Alzheimer’s [168], pulmonary diseases [169, 170], and heat stress [171].
However Sargent’s [172] review of literature comparing ANN with standard statistical techniques concluded that ANN should not replace standard statistical approaches as the method of choice for the classification of medical data. This result was supported by studies in urology [173] and oncology [174]. It is suggested that ANN methods should continue to be used and explored with statistical approaches in a complementary manner. A disadvantage of ANN models has been that they do not provide explanations for their decisions. Hence they are regarded as black-boxes. Recently some progress has been made to address this limitation [175].

Figure 6: Representation of an artificial neural network (ANN) for diagnosis of malaria with symptoms as input features.

3.7 Bayesian Network

Bayesian belief networks, also referred to as probabilistic causal networks or Bayesian Networks (BNs), overcome the assumption of independence between findings which was associated with earlier, NB approach. BNs are a merger of symbolic reasoning and bayesian approaches. The nodes in the network represent the findings and the links interconnecting the nodes model the inter-dependencies of the findings (figure 7).

The early applications of BNs were published in mid-nineties. BNs have been used for radiology [176], oncology [177, 178, 179, 180], asthma [181], penetrating trauma [182], pneumonia [183], pancreatitis [184], cardiology [185], hematology [186], neuro-muscular disorder [187], and hyper-kinetic disorder [188]. QMR-DT, Pathfinder, Promedas, DIAMED and miniTUBA are some systems using bayesian networks. Shwe et al. [189] reformulated QMR/Internist database into a bayesian network, called QMR-DT. This system showed a diagnostic accuracy comparable to the original Internist I, despite the many approximations that were used in the conversion process. Promedas is a commercially available system based on knowledge acquired from medical literature and medical experts [190]. It includes 2000 diagnoses, 1000 findings and 8600 connections between diagnoses and findings, covering a large part of internal medicine. DIAMED has been shown to be useful to reason on medical data [191]. miniTUBA [192] is a web-based modeling system for temporal datasets.

BN models can be complex and practically intractable for some problems. Reducing the complexity of BN and optimization of approximation algorithms for inference, is an active research area [193].

3.8 Support Vector Machines

Support Vector Machine (SVM) methods map data points to a multi-dimensional space and attempt classification by constructing a high dimensional hyperplane to separate the points. One of the earliest report using SVM for diagnoses is by Chan [194], for glaucoma detection by identifying patterns obtained by Standard Automated Perimetry (SAP). Utility of SVMs has been reported for diagnosis of cerebral palsy [195], Parkinsonism [196], cancer [197, 198, 199, 200, 201] and diabetic nephropathy [202]. Work on application of SVM for MDS is at a nascent stage, but with promising results.
3.9 Other

Apart from the major paradigms of development noted above, several other approaches have been investigated for MDS. These include sequential diagnosis [203], cluster analysis [204], discriminant analysis [205, 206, 207], logistic regression [208, 209, 210, 211, 212, 213], partitioning methods [214], Dempster-Schafer theory [215], matrix discrimination [216], modified constellation graph methods [217], degree of similarity [218], nearest neighbor [219], case-based (instance-based or memory-based) learning [220, 221, 222, 223, 224], rough sets [225], decision-tree induction [226, 227, 228], Testor theory [229], information theory [230] and biased min-max probability machine [231].

This list of modeling techniques applied for MDS is not exhaustive, and medical diagnosis is a popular problem for researchers working on classification algorithms. Consequentially, a choice from a wide variety of techniques is available for MDS application development.

4 Discussion

In this section we have attempted to describe the general trends in research on modeling techniques for MDS, and to identify short-comings as well as provide future directions.

4.1 Lack of decision accuracy

MDS systems for small domains like ECG or EEG classification, have become popular but systems for more complex problems have not yet achieved the acceptable degree of accuracy. This continues to be a cardinal
obstacle in the utilization of MDS applications in the clinic. There is a need for research to develop accurate classification algorithms. Apart from the classification/inference algorithm, improved representation of patient information may lead to performance improvement. For instance representing temporal and vague concepts [232, 233] and integrating data from diverse sources [234], as elaborated below have been shown to improve accuracy.

4.2 Semi-automatic to automatic knowledge acquisition

In the 1970s and 80s, many researchers focused on input of physicians to construct the knowledge-base [235, 79] for MDS tools. Over the years, the alternative approach of automated analysis of datasets to construct knowledge-base, has become more popular [236, 103]. The latter approach offers several advantages:

- Knowledge-Bases derived from datasets are more precise in comparison with knowledge-bases constructed from expert inputs. This is because the inputs provided by human experts are often vague, due to limited grades of perception [4, 237].

- When the knowledge-base has a large number of parameters, the process of collating experts’ inputs in the former approach, becomes cumbersome and practically unfeasible [235, 238]. Automatic approach obviates the logistical difficulties.

- Knowledge-Base constructed using the automated approach captures empirical evidence in the data. This approach aligns with the evidence-based decision making movement which emphasizes on use of empirical evidence to make clinical decisions [239].

- Medical datasets embed local epidemiological patterns. Hence the derived knowledge-bases can result in more accurate MDS tools, as disease and symptom patterns vary from one region to another [240, 241, 242, 47]. The physician experts on the other hand may not be aware of local trends, especially when they do not have sufficient experience of clinical practice in a particular locality and when there is a lack of systematic studies of disease patterns for the location.

4.3 Choice of modeling technique

The early MDS systems were based on NB model, which made assumptions of conditional independence between symptoms and mutual exclusivity of diseases. These systems performed as-well-as, and at times better than experts. However, despite their performance NB did not receive wider applications, for the following reasons [243]: 1) the assumptions of NB, were considered as an over-simplification and 2) the quantitative approach of Bayes rule was perceived as inappropriate representation of the qualitative way in which humans reason.

AI techniques which were emerging at that time, appeared to provide alternative solution to bypass these problems. The Confidence Factor (CF) [244] and QMR inference model, used heuristic methods for inference and instead focused on representing large amount of expert knowledge. However, the heuristic inference mechanisms did not avoid the assumptions of conditional independence in NB; they merely obscured them [245]. Theoretical studies have shown that these heuristic schemes and the Dempster Schafer [246] model were related to odds likelihood updating scheme [247, 248]. An additional disadvantage of the AI systems was that the parameters for these systems were more likely to be erroneous as compared to the NB model. This is because the assessment of symptom-disease relationships for these systems required the physicians to provide judgements in terms of likelihood of a disease given a finding in a direction that is considered cognitively challenging [249, 243]. In contrast, the use of Bayes rule required the estimation of probabilities in terms of likelihood of a symptom given a disease, which is natural to physicians. In an empirical study, Heckerman [250] found that NB had better diagnostic accuracy than CF.

These developments led researchers in 1990s to revert to the principled probability based approach for inference, but now coupling it with expressive and richer knowledge representation schemes, which led to the development of BNs. Other methods like SVM, ANN and Information theory have also been explored. However, despite the development of new complex models [174], no significant improvement in accuracy has been achieved and many comparative studies have evaluated NB as near optimal [251, 39, 40]. NB remains the most widely
researched approach and many researchers regard NB as the reference standard and use it as a benchmark for performance of any new algorithm. The soft computing methods including FST, Evolutionary Algorithm (EA), ANN, and BN are computationally intensive. These methods have benefitted from the availability of cheap computational power that has rendered the feasibility of some application, earlier regarded as intractable.

Overall a wide choice of modeling techniques is available for developers of MDS tools. However, there is a dearth of studies which compare the performance of different techniques [252]. Hence, we suggest that a diversity of techniques are evaluated for the problem under consideration, in order to decide on the optimal technique.

4.4 Integrating clinical and genomic patient information

One of the reasons, why MDS tools may be found to have sub-optimal accuracy is that the training data may itself lack all the attributes that are required for decision making. Combining decision support methodologies that process information stored in different data formats has been shown to improve MDS performance [253]. Apart from laboratory information, attributes extracted from free text information, and even gene profiling data could be combined in the Electronic Medical Record (EMR). Decision models built from such an integrated EMR may possibly lead to clinically applicable systems. However gene profiles are complex and analysis of such data is currently limited to bioinformatics applications [254].

4.5 Modeling temporal information

Knowledge about the order of occurrence of events, has a significant bearing on clinical decision. For example, head injury may occur after fainting or fainting may occur after head injury – these situations differ in the causative factor and consequently the diagnosis. Currently, most MDS systems are not designed with a formal representation of time, and they are unable to reason about temporal relations of events. In the last decade, some temporal models [255, 256] have been developed and there exists empirical evidence that temporal modeling can improve MDS accuracy [257, 145]. Temporal modeling of medical information is challenging and may hold the key for improving performance of MDS systems. The reader is referred to [258, 259, 260] for further reading.

4.6 Modeling fuzziness

Probability based approaches have dominated MDS research. However a problem with probabilistic approaches is that they do not differentiate between the types of uncertainties, and model all uncertainty using probability theory [15]. Uncertainty in the entities involved in medical decision making can be differentiated into ignorance, non-specificity, vagueness and strife [261]. FST provides a promising alternative, as it offers both – a mechanism to model the different types of uncertainties as well as a principled approach for inference using fuzzy entities. Several studies have demonstrated the utility of fuzzy logic concepts in medical domain (section 3.5). FST has been recently of active interest for MDS research.

FST can effectively model medical reasoning because, symptom and disease entities in medicine are defined using linguistic descriptions which are inherently vague/fuzzy. The imprecise definitions along with the human cognitive limitation render the physicians reasoning process vague and subjective. Although models based on FST generally provide a less precise representation of a problem, as compared to models based on Aristotelian logic (including probability based models), FST based models have been known to closely approximate human intelligence [261]. Hence, the FST models could provide a better approximation of physician’s diagnostic process, which is characterized by the lack of precision in knowledge and an inference mechanism, which has eluded attempts of probabilistic modelers.

4.7 Evaluation measures

MDS systems are often designed to output differential diagnosis. This is essentially a list of decision classes in the decreasing order of their confidence scores. Medical diagnostic problems often involve more than two decision classes (referred to as multi-class problem), which complicates the computation of decision accuracy of the differential diagnosis. This is because much of the work on evaluation measures for classification algorithms,
has been carried out for problems in which the decision attribute is binary—there are only two decision classes. Most of the evaluation studies carried out before the last two decades, measured performance by identifying the rank of the correct diagnosis in the list of differential diagnosis—a lower rank meant better accuracy. A cut-off rank was used, and the percentage of times the correct diagnosis was present below the cut-off rank, was reported as the accuracy. However, this measure is subject to the number of diagnoses considered in the system and also does reflect the sensitivity and specificity of the system. To address this problem, some researchers started reporting area under Receiver Operating Characteristic (ROC) [202], for binary classification problems. Recently measures generalizing the computation of ROC for multi-class problems have also been developed [263]. However, researchers continue to report evaluation results using different measures, which impedes the task of comparing the algorithms. Further research is needed to establish standard evaluation measures for reporting performance of Medical Decision Support System (MDS) applications.

4.8 Evaluation in the clinical setting

MDS models are usually evaluated on small or simulated datasets for preliminary studies, followed by comprehensive evaluations on medium to large (> 200) [172] real datasets (see figure 9). Few evaluation studies are performed on large datasets [236, 264]. Reliability of the results generally increases with size of the dataset and number of evaluative studies. Evaluations are usually retrospective, in which the case records of previously labelled patients are used as gold standard. Rarely, MDS models are evaluated prospectively in the clinical setting, where the model’s decision are compared with those of the physician in real-time. The latter evaluations are most useful to build confidence about computer-based decision aids in the medical community. MDS researchers should perform retrospective evaluations first, and when their algorithms are found to be sufficiently accurate, should aim for prospective studies [150, 265]. De Dombal pioneered clinical evaluations of MDS systems. His Leeds system was evaluated first retrospectively on past patient records, and then with real-time prospective studies across many locations involving practicing surgeons [38]. This research played a significant role in impressing the practical utility of MDS systems.

![Histogram of the number of data instances in small and medium-large datasets investigated in the surveyed literature](image)

**Figure 9:** Histogram of the number of data instances in small and medium-large datasets investigated in the surveyed literature

4.9 Complexity of decision problems

The complexity of a decision problem depends on i) domain, ii) number of instances in training data, iii) number and type of attributes, and iv) number of diagnostic labels considered. Medical domains vary in the degree of complexity of the classification problem they deal with. For instance, cardiovascular diseases which occur with numerous symptoms, present a low complexity problem as compared to cancers which are insidious with few
symptoms. In addition, a decision problem may be complex due to dearth of training instances and information attributes. Binary diagnostic problems in which the presence or absence of a disease is to be computed, are less complex than problems in which patients can have one or more diagnosis from a diagnosis set. The problem domains addressed in MDS research are diverse (see table 1 and figures 10 and 9). Researchers should evaluate simpler problems for preliminary studies and should progress towards increasingly complex problems to explore the utility of their algorithms.

![Pie chart showing distribution of studies cited in this paper, across medical specialities.](image)

**Figure 10**: Distribution of studies cited in this paper, across medical specialities.

### 4.10 Decision Utility

After arriving at a provisional diagnosis based on patient interview, physicians consider severity of the differential diagnoses and individual patient preferences, to calculate the utility of diagnoses or expected utility of prescribing additional tests [239, 78]. Utility considerations involve the risk to the patient and cost of the tests (see section 3.2). Physicians have a tendency to give a guarded opinion [280], wherein they do not easily discard diagnosis, which poses potential risks to the patient. The lack of such utility considerations, could be a reason why evaluations comparing the decisions of MDS tools with physician’s differential diagnosis, report low accuracy. This factor is expected to play a lesser role, when the final diagnosis arrived at discharge of the patient from the health care facility, is considered as the gold standard for measuring performance accuracy. Hence, research is needed to investigate the significance of decision utility for MDS evaluation experiments.

### 4.11 Utilization of High Performance Computing

Technological advances in the field of high performance [281] computing have made cheap computational power available to individuals. Multi-core processors and computers in parallel configurations have been constructed for high end computational applications. This provides an opportunity for exploring MDS modeling techniques that were regarded practically intractable, a few decades ago. MDS researchers should harness these resources, so developed MDS tools for increasingly complex medical decision problems.

### 4.12 Opensource

Many MDS systems are developed for commercial applications. Their decision algorithms or knowledge-bases are rarely released into public domain. The medical datasets used for evaluations are usually withheld. Non-availability of the tools, knowledge-bases and data impedes replication of the evaluation experiments as well as comparative studies. If researchers were to release these materials in opensource, it may be possible to reuse them in evaluation experiments of other algorithms, and useful benchmarks can be obtained. The importance of the need to release the tools and datasets as supplementary material with the reports, was realized early on by researchers in bioinformatics and genomic domain. Consolidation of their efforts has fostered advances in those fields. A similar practice needs to be encouraged in MDS research. Another reason why MDS tools need to be opensource is their potential application in the clinical setting, where error-free operation is called for [282].
Table 1: Diagnostic problems addressed in surveyed literature

<table>
<thead>
<tr>
<th>Specialty</th>
<th>Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiovascular</td>
<td>myocardial-infarction [150, 69, 208, 152, 210], coronary disease [215, 125], arrhythmia [104], cardiac disease [113, 121, 185, 231, 102, 132, 154], giant cell arteritis [153], hypertension [68], pulmonary embolism [266], valvular disease [155]</td>
</tr>
<tr>
<td>ENT</td>
<td>otoneurological disease [227], apnea [103], sore throat [139]</td>
</tr>
<tr>
<td>Gastroenterology</td>
<td>hepatobiliary [127, 216, 93], appendicitis [166, 37, 41, 37], bowel disease [71], obstruction [72], cirrhosis [267, 49], dyspepsia [163, 220], hepatitis [146, 123, 47, 51, 48, 50], bleeding [73, 133], pain in abdomen [123, 38, 27, 251, 40, 42, 55], pancreatitis [184, 119, 53], ascites [87],</td>
</tr>
<tr>
<td>Neurology</td>
<td>Alzheimer’s [168], Parkinsonism [196], acute event [94], aphasia [124], cerebral palsy [195], epilepsy [60], headache/facial pain [90], general [268], neuromuscular disorders [187, 269]</td>
</tr>
<tr>
<td>Oncology</td>
<td>cancer of breast [161, 161, 179, 142, 146, 123, 231, 222, 198, 201], cerebrum [46], cervix [270, 271, 178], esophagus [84], stomach [164], ovary [197], prostate [173], skin [162, 83], lung [45, 54]; leukemia [159, 160, 141], melanoma [199], neuroma [44], general [43, 200]</td>
</tr>
<tr>
<td>Psychiatry</td>
<td>ADHD [188], geriatric [111], general [157, 206, 85, 86]</td>
</tr>
<tr>
<td>Reproductive</td>
<td>contraception [123], hysterectomy [101], gynecological pain in abdomen [39], pregnancy disorders [89], vaginal discharge [59]</td>
</tr>
<tr>
<td>Respiratory</td>
<td>asthma [181, 212], interstitial disease [32, 209], lung disease [98, 169], pneumonia [169, 170, 183, 134], pleural effusion [75],</td>
</tr>
<tr>
<td>Endocrinology</td>
<td>diabetes [146, 224], thyroid disorder [272, 146, 99], general [268]</td>
</tr>
<tr>
<td>Hematology</td>
<td>anaemia [138, 67], blood donation [123], general [186, 36, 66, 273]</td>
</tr>
<tr>
<td>Immunology</td>
<td>rheumatic disease [134, 118, 107, 112, 61, 62, 221, 82]</td>
</tr>
<tr>
<td>Infectious</td>
<td>Lyme’s [167], MSRA [165], malaria [128], post-operative infection [126], nosocomial infection [92]</td>
</tr>
<tr>
<td>Musculoskeletal</td>
<td>low-back pain [158], arthritis [122], intrabony lesion [57], osteoporosis [135]</td>
</tr>
<tr>
<td>Opthalmology</td>
<td>glaucoma [81, 194], leukocoria [65], general [268]</td>
</tr>
<tr>
<td>Other</td>
<td>penetrating trauma [182], pediatric [230], critical care [274] disease outbreak [31], triage [275], pulpal pain [56], renal [90, 276], urinary incontinence [88], pathology [277, 144], general [190, 123, 110, 17, 28, 80, 191, 25], radiology [176, 177, 278, 279], heat-stress [171], metabolic disorder [91], congenital anomaly [33, 34, 114], adverse drug reaction [70], lymph-node [78], rare syndromes [220], skin disease [64, 63, 146, 131]</td>
</tr>
</tbody>
</table>
4.13 Using free-text data

A large proportion of patient information is in the free text form. This is mainly comprised of a verbatim report of patient chief-complaints [283] and physicians’ notes. Such data lacks structure and is not amenable to analysis. Hence, it is largely unused by decision support tools. For extracting information attributes, free-text is parsed by NLP algorithms to identify the parts of speech and these are mapped to a standard vocabulary (ontology). This area of research is rapidly progressing [284] and applications [285, 21, 286, 287] based on free-text parsing have been developed. With advances in this area, free-text can provide useful information attributes, and integration of these attributes with laboratory-test based ones will enhance the decision accuracy of MDS applications.

Unified Medical Language System (UMLS) [288] reflects the progress in representing complex medical concepts using ontologies. Advances in ontology and parsing would hugely facilitate the extraction of features from free text data, that was not earlier amenable to analysis. All these developments, along with the increased use of Electronic Medical Record (EMR) systems in hospitals, can be expected to lead to the availability of huge amounts of data features. MDS applications would be needed to harness the intelligence derived from data in the EMR systems.

5 Concluding Remarks

A primary goal of biomedical informatics is to provide computer based decision aids to physicians. Research for developing increasingly accurate MDS algorithms is necessary to achieve this goal. The need for MDS can be expected to become increasingly acute, as information explosion in medical science continues. Implementation of EMR systems at health-care centers can facilitate adoption of MDS applications.

The authors suggest that – i) Improvement in the accuracy of MDS application may be possible by modeling of vague and temporal data, research on inference algorithms, integration of patient information from diverse sources and improvement in gene profiling algorithms; ii) MDS research would be facilitated by public release of de-identified medical datasets, and development of opensource data-mining tool kits; iii) Comparative evaluations of different modeling techniques are required to understand characteristics of the techniques and to guide developers in the choice of technique for a particular medical decision problem; iv) Evaluations of MDS applications in clinical setting are necessary to foster physicians’ utilization of these decision aids.

Acronyms

AI Artificial Intelligence.
ANN Artificial Neural Network.
BN Bayesian Network.
CF Confidence Factor.
EA Evolutionary Algorithm.
ECG Electrocardiogram.
EMR Electronic Medical Record.
FR fuzzy relation.
FST fuzzy set theory.
MDS Medical Decision Support System.
MDS Medical Decision Support.

NB naive bayes.

NLP Natural Language Processing.

QMR Quick Medical Reference.

ROC Receiver Operating Characteristic.

SAP Standard Automated Perimetry.

SVM Support Vector Machine.

UMLS Unified Medical Language System.

6 Acknowledgments

The authors are grateful for the excellent suggestions and comments of the reviewers that helped improve the manuscript. The first author was supported by a fellowship from Indian Council of Medical Research (ICMR), when part of this research was carried out.

References


Enzo Grossi, Massimo P. Buscema, David Snowdon, and Piero Antuono. Neuropathological findings processed by artificial neural networks (anns) can perfectly distinguish alzheimer’s patients from controls in the nun study. *BMC neurology*, 7:15+, June 2007.


