Chapter 11
Diagnostic Decision Support Systems

Randolph A. Miller

Abstract Since primeval times, mankind has attempted to explain natural phenomena using models. For the past five decades, a new kind of modeler, the healthcare informatician, has developed and proliferated a new kind of model, the clinical Diagnostic Decision Support System (DDSS). This chapter presents a definition of clinical diagnosis and of DDSS; a discussion of how humans accomplish diagnosis; a survey of previous attempts to develop computer-based clinical diagnostic tools; a discussion of the problems encountered in developing, implementing, evaluating, and maintaining clinical diagnostic decision support systems; and a discussion of current and future systems.

Keywords Diagnostic decision support • Diagnosis • Human reasoning • Evaluation • Decision support system development

Since primeval times, mankind has attempted to explain natural phenomena using models. For the past five decades, a new kind of modeler, the healthcare informatician, has developed and proliferated a new kind of model, the clinical Diagnostic Decision Support System (DDSS). Modeling was historically, and still remains, an inexact science. Ptolemy, in the Almagest, placed the earth at the center of the universe and still could explain why the sun would rise in the east each morning. Newton's nonrelativistic formulation of the laws of mechanics works well for earthbound engineering applications. Yet mankind, using imperfect models, has built machines that fly, and has cured many diseases. Past and present DDSS incorporate

---

Portions of this chapter have been taken verbatim, with permission of the American Medical Informatics Association (AMIA), which owns the copyrights, from: Miller RA, Medical Diagnostic Decision Support Systems—Past, Present, and Future: A Threaded Bibliography and Commentary. J Am Med Inform Assoc 1994;1:8–27; and from Miller RA, Evaluating Evaluations of Medical Diagnostic Systems, J Am Medical Inform Assoc 1996;3:429–431. Dr. Miller acknowledges the earlier contributions of Antoine Geissbuhler, MD, of the Hospital of the University of Geneva, Switzerland, who co-authored a previous version of this chapter. The author thanks Joyce Green for her assistance with copy editing.

R.A. Miller, M.D. (✉)
Department of Biomedical Informatics, Vanderbilt University Medical Center,
2525 West End Avenue, Ste 1475, Nashville, TN 37203, USA
e-mail: randolph.a.miller@vanderbilt.edu

E.S. Berner (ed.), *Clinical Decision Support Systems*, Health Informatics,
DOI 10.1007/978-3-319-31913-1_11

inexact models of the incompletely understood and exceptionally complex process of clinical diagnosis. Because DDSS augment the natural capabilities of human diagnosticians, they have the potential to be employed productively [1].

This chapter presents a definition of clinical diagnosis and of DDSS; a discussion of how humans accomplish diagnosis; a survey of previous attempts to develop computer-based clinical diagnostic tools; a discussion of the problems encountered in developing, implementing, evaluating, and maintaining clinical diagnostic decision support systems; and a discussion of current and future systems.

11.1 Definitions of Diagnosis

To understand the history of clinical diagnostic decision support systems and envision their future roles, one must first define clinical diagnosis and computer-assisted clinical diagnosis. A simple definition of diagnosis is: [2]

> *the placing of an interpretive, higher level label on a set of raw, more primitive observations* [Definition 1].

By this definition one form of diagnosis might consist of labeling as "abnormal" any laboratory test results falling outside 1.5 times the 95 % confidence intervals for the "normal" values seen in the general population as measured by that laboratory. Another level of diagnosis under the same definition might consist of labeling the combination of a low serum bicarbonate level, a high serum chloride level, and an arterial blood pH of 7.3 as "metabolic acidosis." A more involved definition of diagnosis, specific for clinical diagnosis, is: [2]

> *a mapping from a patient's data (normal and abnormal history, physical examination, and laboratory data) to a nosology of disease states* [Definition 2].

Both of these definitions treat diagnosis improperly as a single event, rather than as a process. A more accurate definition appeared in the Random House Collegiate Dictionary: [3]

> *the process of determining by examination the nature and circumstances of a diseased condition* [Definition 3].

Skilled diagnosticians develop an understanding of what the patient's life situation was like before the illness began, how the illness has manifested itself, and how it has affected the life situation [2]. The clinician must also determine the patient's understanding of, and response to, an illness. The process of diagnosis entails a sequence of interdependent, often highly individualized tasks: evoking the patient's initial history and physical examination findings; integration of the data into plausible scenarios regarding known disease processes; evaluating and refining diagnostic hypotheses through selective elicitation of additional patient information, such as laboratory tests or serial examinations; initiating therapy at appropriate points in time (including before a diagnosis is established); and evaluating the effect of both the illness and the therapy, on the patient, over time [2].

Diagnosis is a process composed of individual steps. These steps go from a point of origin (a question and a set of "presenting findings" and "previously established diagnoses"), to a point of destination (an answer, usually consisting of a set of "new established diagnoses" and/or "unresolved differential diagnoses"). While the beginning and end points may be identical, the steps one diagnostician follows may be very different from those taken by another diagnostician, and the same diagnostician may take different steps in two nearly identical cases. Because expertise varies among clinicians, different individuals will encounter different diagnostic problems in evaluating the same patient. For instance, they may experience dissimilar difficulties at disparate steps in the diagnostic process, even if they follow exactly the same steps.

Studies of clinicians' information needs help us to understand the variability in diagnostic problem solving among clinicians. Osheroff and colleagues [4, 5] used participant observation, a standard anthropological technique, to identify and classify information needs during the practice of medicine in an academic health center. They identified three components of "comprehensive information needs:" (1) currently satisfied information needs (information recognized as relevant to a question and already known to the clinician); (2) consciously recognized information needs (information recognized by the clinician as important to know to solve the problem, but which is not known by the clinician); and (3) unrecognized information needs (information that is important for the clinician to know to solve a problem at hand, but is not recognized as being important by the clinician). Failure to detect a diagnostic problem at all would fall into the latter category. Different clinicians will experience different diagnostic problems within the same patient case, based on each clinician's varying knowledge of the patient and unique personal store of general medical knowledge. Osheroff et al. noted the difficulty people and machines have in tailoring general medical knowledge to specific clinical cases. There may be a wealth of information in a patient's inpatient and outpatient records, and also a large medical literature describing causes of the patient's problems. The challenge is to quickly and efficiently reconcile one body of information with the other [1, 4].

A DDSS can potentially facilitate that reconciliation. A DDSS can be defined as:

> *a computer-based algorithm that assists a clinician with one or more component steps of the diagnostic process* [Definition 4].

While individual clinicians attach different meanings to "diagnosis", users of DDSS are often slow to recognize that each system functionally defines diagnosis as the set of tasks that the DDSS can perform. Experienced users employ DDSS as tools to supplement, rather than replace, their own diagnostic capabilities. Naive users view diagnosis on their own terms, based on their own experiences, and expect DDSS to behave in accordance with their assumptions. Untrained DDSS users' unrealistic, preconceived expectations can engender subsequent frustrations. For example, a DDSS cannot solve a vague problem with minimal input; nor is a DDSS likely to help in understanding how an illness has affected a patient's lifestyle. Conversely, system developers sometimes create useful diagnostic tools that provide capabilities outside the experience of everyday clinical practice. For example,

the relationships function of R-QMR[1] (a DDSS), takes, as input, up to ten findings that the clinician-user would like to explain as the key or "pivotal" findings from a diagnostically challenging case, and produces, as output, a rank-ordered list of "disease complexes" that each explain all of the input findings [7]. Each disease complex is made up of from one to four interrelated disorders (e.g., disease A predisposing to disease B and causing disease C). Because busy clinicians can spare little free time for extraneous activities, user training for DDSS utilization is extremely critical and must address the potential cognitive mismatch between user expectations and system capabilities.

That the problem to be solved originates in the mind of the clinician-user is conceptually critical for DDSS development, implementation, and evaluation. The diagnostic problem cannot be defined in an absolute sense, for example, by presenting an arbitrary set of input findings selected from a case—i.e., if clinical findings are extracted from a patient case in the absence of a query from a clinician caring for the patient, do those findings comprise a diagnostic problem to be solved? In only one situation can the findings of a case, in isolation, define a diagnostic problem: when the diagnostic problem is the global one. In the global problem, the DDSS, through its own initiative, takes all the steps in the diagnostic process required to explain all patient findings, by "concluding" new diagnoses (or listing unresolved differential diagnoses if no solution exists). Practicing clinicians rarely encounter the "global" diagnostic problem. Healthcare providers usually complete a portion of the diagnostic evaluation process before they encounter difficulty in making a diagnosis, and, correspondingly, once they overcome the difficulty (e.g., by consulting a colleague), they are usually capable of completing the evaluation without further assistance. While early DDSS developers often assumed the only problem worth solving was the global diagnostic problem, emphasis over the last decades has shifted to helping clinicians with problems they encounter during individual steps in the diagnostic process. This has led to the demise of the "Greek Oracle" model, wherein the DDSS was expected to take all of the patient's findings and come up with "the answer"[8]. Current DDSS models assume that the user will interact with the DDSS in an iterative fashion, selectively entering patient information and using the DDSS output to assist with the problems that the user has encountered in the diagnostic process [9].

To interact optimally with a DDSS, users must understand assumptions built into the system. Each DDSS functionally defines diagnosis as the tasks it can perform (or assist users in performing). The subtle nature of underlying assumptions incorporated into DDSS can be deceptive. As an example, one of the most well-known diagnostic systems is the Bayesian program for diagnosis of acute abdomi-

[1] In this chapter, R-QMR refers to the noncommercial, research version of QMR, the DDSS developed by Miller et al. [6]. The commercial version of QMR, previously marketed by First DataBank, while initially identical to R-QMR in 1990, was developed independently of R-QMR after that time. The commercial version of QMR is no longer marketed. Since 2014 Miller and colleagues at Vanderbilt have been developing a third-generation non-commercial successor system, "AskVanderbilt".

nal pain developed by de Dombal and colleagues [10, 11]. The system's original goal, not stated explicitly, was to discriminate among surgical and nonsurgical causes of acute abdominal pain in an emergency room (or similar) setting. The system supported a limited number of explicit diagnoses; all except "nonspecific abdominal pain," were potentially surgical conditions (such as acute appendicitis, acute pancreatitis, and acute diverticulitis). The performance of the system was evaluated in multicenter studies [11] and shown to be exemplary with respect to the circumstances for which it was designed. Nevertheless, de Dombal's system would most likely disappoint naive users relying on it to diagnose patients presenting with acute abdominal pain in more general settings. The system could not properly diagnose patients presenting with acute intermittent porphyria, lead poisoning, early T10 dermatome herpes zoster, or familial Mediterranean fever. The system would correctly label those conditions as "nonspecific abdominal pain," even though some are potentially life threatening and treatable. Clinical users of DDSS in general should recognize the potential for errors when using DDSS. This mandates that clinicians supplement DDSS-based suggestions with their own expert knowledge.

The utility of making specific diagnoses lies in the selection of effective therapies, making accurate prognoses, and providing detailed explanations [1]. In some situations, it is not necessary to arrive at an exact diagnosis in order to fulfill one or more of these objectives. Treatment is often initiated before an exact diagnosis is made (e.g., patients in the emergency room receive oxygen for shortness of breath, before the etiology is known). Furthermore, the utility of making certain diagnoses is debatable, especially if there is a small probability of effective treatment.

The cost of eliciting all possible patient data is potentially staggering—temporally, economically, and ethically—since there are real risks of morbidity and/or mortality associated with many diagnostic procedures such as liver biopsy or cardiac catheterization. Given the impossibility and impracticality of gathering every conceivable piece of diagnostic information with respect to each patient, the "art" of diagnosis lies in the ability of the diagnostician to carefully evoke enough relevant information to justify all important and ultimately correct diagnoses in each case, as well as to initiate therapies at appropriate points during the evaluation [2].

The knowledge of how to "work up" the patient depends critically on the ability to evoke history, symptoms, and physical examination findings, concurrently with the ability to generate diagnostic hypotheses that suggest how to further refine or pursue the findings already elicited, or to pursue completely different additional findings. In addition, this must be done in a compassionate and cost-effective manner [2].

11.2 Human Diagnostic Reasoning

Diagnostic reasoning involves diverse cognitive activities, including information gathering, pattern recognition, problem solving, decision-making, judgment under uncertainty, and empathy. Large amounts of highly organized knowledge are

necessary to function in this relatively unstructured cognitive domain. Our knowledge of human diagnostic reasoning is based on generic psychological experiments about reasoning and on direct studies of the diagnostic process itself. Relevant principles of human problem-solving behavior have been unveiled through focused studies examining constrained problem spaces such as chess-playing and cryptarithmetic [12]. Such studies have documented that experts recognize patterns of activity within a domain at an integrated, higher level ("chunking") than novices. Additional psychological experiments about judgments made under uncertainty [13] have provided insights into individuals' imperfect semi-quantitative reasoning skills.

To investigate the complex intellectual task of clinical diagnosis, many researchers [14, 15] have used behavioral methods that combine protocol analysis with introspection. Researchers record clinicians as they think aloud while performing specified cognitive tasks related to diagnosis (including normal clinical activities). Post facto, the clinicians themselves, or others, are asked to interpret the motives, knowledge, diagnostic hypotheses, and strategies involved in the recorded sessions. However, there is no proof that the stories constructed by experts to explain their diagnostic reasoning correspond to the actual reasoning methods they use subconsciously.

Most models of diagnostic reasoning include the following elements: the activation of working hypotheses; the testing of these hypotheses; the acquisition and interpretation of additional information; and confirming, rejecting, or adding of new hypotheses as information is gathered over time. Working hypotheses are generated early in the process of information gathering, at a time when only few facts are known about the patient [14, 15]. Only a limited number of these hypotheses, rarely more than five, are entertained simultaneously, probably due to the limited capacity of human short term memory [16]. Early hypothesis generation is accomplished through some form of pattern recognition, with experts more capable of applying compiled knowledge and experiences than novices. Comparing clinical reasoning in novices and experts, Evans and Patel [17] showed that experts rarely rely directly on causal reasoning and knowledge of basic sciences, except when reasoning outside their domain of expertise.

As noted by Pople and others [8], clinical diagnosis fits Nobel Laureate Herbert Simon's criteria for being an ill-structured problem [18]. Simon gave as an example of an ill-structured problem, the task an architect faces in creatively designing a new house "from scratch"—the realm of possible solutions encompasses a great variety of applicable methods and a broad set of alternative outcomes. As noted by Pople, Simon observed that one can solve ill-structured problems by splitting the problems into smaller, well defined subtasks that are each more easily accomplished [8].

In clinical diagnosis, early hypothesis generation helps to constrain reasoning to "high yield" areas, and permits the use of heuristic methods to further elucidate a solution [19]. Studies have shown that most clinicians employ the hypothetico-deductive method after early hypothesis generation [14, 15]. Data are collected with a view to their usefulness in refining, rejecting, or substituting for the original set of hypotheses. In the setting of clinicopathological exercises, Eddy and Clanton [20]

showed that identification of a pivotal finding is often used to simplify the diagnostic problem and to narrow the focus to a limited set of hypotheses. Kassirer and Gorry [15] described the "process of case building," where hypotheses are evaluated against the model of a disease entity using techniques that can be emulated in computers using Bayes' rule, Boolean algebra, or template matching (see Chap. 2 for an explanation of these terms). They also recognized that heuristic methods are commonly used to confirm, eliminate, discriminate between, or explore hypotheses. Weed [21] and later Hurst and Walker [22] suggested that clinical problem solving can be approached by splitting complex, composite problems into relatively independent, discrete "problem areas." With respect to diagnosis, Pople (like Gorry earlier) observed that separating complex differential diagnoses into problem areas allows diagnosticians to apply additional powerful reasoning heuristics. They can assume that the differential diagnosis list within a problem area that contains mutually exclusive hypotheses and that the list can be made to be exhaustive (i.e., complete), so that it is assured that the correct diagnosis is on the list for the problem area, and that only one diagnosis on the list is the correct one [8].

Kassirer identified three abstract categories of human diagnostic reasoning: probabilistic, causal, and deterministic [23]. Formal models for each type of reasoning have been developed—at times independently of observational studies on how actual reasoning occurs. Approaches such as Brunswik's lens model [24], Bayesian algorithms [25, 26], and decision analysis [27, 28] define statistical associations between clinical variables and use formal mathematical models to derive "optimal" decisions. While diagnosticians clearly consider prevalence and other likelihood-related concepts during their reasoning [14, 15], observational and experimental studies show that clinicians do not calculate probabilities subconsciously during their own diagnostic reasoning [13, 29]. Human problem solvers tend to rely on judgmental heuristics. Experiments document that humans improperly evaluate subjective probabilities, misuse prior probabilities, and fail to recognize important phenomena, such as the regression towards the mean.

Evidence indicates that humans have more difficulty reasoning with probabilities than they do understanding the concepts that underlie them [30]. Humans also fall prey to reasoning errors such as reluctance to revise opinions when new data do not fit with working hypotheses, even when the data's diagnostic significance is properly understood [13, 29].

Models of causal (pathophysiological) reasoning, such as those developed by Feinstein [31, 32] in the 1970s, establish cause-and-effect relations between clinical variables within anatomic, physiologic, cellular, molecular, and biochemical representations of the reality. Although causal inferences (deductive reasoning from causes to consequences) can be viewed as the inverse of diagnostic inferences (abductive reasoning from consequences to causes), studies have shown that when making judgments under uncertainty, humans assign greater impact to causal relationships over other forms of diagnostic data of equal informative weight. Subjects commonly make overconfident predictions when dealing with highly uncertain models [13]. Causal (pathophysiological) reasoning uses shared, global, patient-independent knowledge [32] and provides an efficient means of verifying and

explaining diagnostic hypotheses. Nevertheless, how much causal reasoning is actually used in early hypothesis generation and other stages of non-verbalized diagnostic reasoning is unclear; simple pattern recognition is far more prevalent. Previous studies indicate that experts tend to employ causal, pathophysiological reasoning only when: (a) faced with problems outside the realm of their expertise; (b) solving highly atypical problems, or (c) when they are asked to explain their reasoning to others [5].

In deterministic models, production rules, i.e., specifying appropriate actions in response to certain conditions, are used to represent the basic building blocks of human problem-solving. Such if—then rules representing compiled knowledge can also be expressed in the form of branching-logic flowcharts and clinical algorithms for non-experts to follow. However, production rules do not deal effectively with uncertainty [33], which is a disadvantage in clinical practice, where uncertainty is a common feature.

The late M. Scott Blois, a great philosopher-informatician-clinician, used a funnel to illustrate the spectrum of clinical judgment [34]. Consideration of patients' ill-structured problems, including undifferentiated concerns and vague complaints, occurs at the wide end of the funnel. Focused decisions in response to specific clinical questions (e.g., choosing an antibiotic to treat the exact bacterial species isolated as the cause of a pneumonia) were represented at the narrow end. This model is consistent with Simon's view of how humans solve ill-structured problems [18].

Blois noted that decision support systems were best applied toward the narrow end of the funnel, since circumscribed, well-structured problems are encountered there. Those problems are more amenable to solution through application of computational models of cognitive skills, requiring only focused and specific knowledge. On the other hand, at the open end of the funnel, one has to deal with common-sense knowledge and the general scope of ordinary human judgment in order to make meaningful progress, and few computer-based systems (other than those for record-keeping) are applicable.

11.3 Historical Survey of Diagnostic Decision Support Systems

The literature prior to 1976 described a majority of the important concepts still relevant to current DDSS development. In a comprehensive 1979 review of reasoning strategies employed by early DDSS, Shortliffe, Buchanan, and Feigenbaum identified the following classes of DDSS: clinical algorithms, clinical databanks that include analytical functions, mathematical pathophysiological models, pattern recognition systems, Bayesian statistical systems, decision-analytical systems, and symbolic reasoning (sometimes called "expert" systems) [35]. This section, without being comprehensive, will describe how some of the early pioneering efforts led to many classes of systems present today.

The many types of DDSS result from the large number of clinical domains to which diagnostic reasoning can be applied, from the multiple steps of diagnostic reasoning described above, and from the variety of difficulties that diagnosticians may encounter at each step. Health care informaticians encountering the term "clinical diagnostic decision-support systems" think primarily of general-purpose, broad-spectrum consultation systems [1].

A useful dichotomy separates DDSS into systems for general diagnosis (no matter how broad or narrow their application domains), and systems for diagnosis in specialized domains such as interpretation of ECG tracings [36]. The general notion of DDSS conveyed in the biomedical literature sometimes overlooks specialized, focused, yet highly successful medical device-associated diagnostic systems. Some simple DDSS help to interpret blood gas results, or assist in categorizing diagnostic possibilities based on the output of serum protein electrophoresis devices, or aid in the interpretation of standardized pulmonary function tests. DDSS for cytological recognition and classification have found successful application in devices such as automated differential blood count analyzers and systems to analyze Papanicolaou smears [1]. Small, focused DDSS are the most widely used form of diagnostic decision support programs, and their use will grow as they are coupled with other automated medical devices [1].

In their classic 1959 Science paper, Ledley and Lusted [25] observed that physicians have an imperfect knowledge of how they solve diagnostic problems. Ledley and Lusted stated that both logic (as embodied in set theory and Boolean algebra) and probabilistic reasoning (as embodied in Bayes' rule) were essential components of medical reasoning. They mentioned the importance of protocol analysis in understanding human diagnostic reasoning. They stated that they had examined how physicians solve New England Journal of Medicine CPC (clinicopathological conference) cases as the foundation for their work on diagnostic computer systems. Their insights provided the basis for work on Bayesian and decision-analytic diagnostic systems carried out over subsequent decades. Both for practical reasons and for philosophical reasons, much work on DDSS has focused on the differences between logical deductive systems and probabilistic systems. Chapter 2 describes these approaches in more detail. What follows is a description of how DDSS have embodied varied reasoning principles.

Logical systems, based on "discriminating questions" to distinguish among mutually exclusive alternatives, have played an important role since the pioneering work by Bleich and his colleagues [37] on acid base and electrolyte disorders. To this day, such systems are applicable to narrow domains, especially those where it is fairly certain that only one disorder is present. When users of a branching logic system incorrectly answer one of the questions posed by the system, they may find themselves "out on a limb" with no way to recover except by starting over from the beginning; the likelihood of such problems increases when multiple independent disease processes interact in the patient. Thus, ideal application areas are those where detailed knowledge of pathophysiology or extensive epidemiological data make it possible to identify parameters useful for dividing diagnostic sets into non-intersecting subsets, based on specific characteristics.

Bayes' rule is applicable to many clinical domains. Following Ledley and Lusted's 1959 publication [25], Warner and colleagues developed one of the first medical application systems based on Bayes' rule. In a 1961 JAMA paper [26], Warner et al. described a Bayesian DDSS for the diagnosis of congenital heart diseases. It utilized probabilities obtained from literature review, from their own series of over 1,000 cases, and from experts' estimates based on self-knowledge of pathophysiology. They emphasized that straightforward application of Bayes' theorem requires independence among the diagnoses and among the findings encompassed in the DDSS. They proposed a method for eliminating the influence of redundant findings. Warner et al. observed how diagnostic systems can easily fail due to false positive case findings and due to errors in the system's database. In their evaluation of their system's performance, they pointed out the need for an independent "gold standard" against which evaluators can judge the performance of the system. For that purpose, they used cardiac catheterization data and/or anatomical (postmortem) data excluded from the inputted case descriptions to confirm the actual patient diagnoses. Warner et al. continued to develop and refine models for Bayesian diagnosis over the years [1]. In 1968, Gorry and Barnett developed a model for sequential Bayesian diagnosis that extended Warner's earlier approach [38].

Many regard the system for the diagnosis of acute abdominal pain developed by de Dombal and colleagues at the University of Leeds as the first practical Bayesian system. It was utilized at widespread clinical sites [1, 10]. A large number of groups have subsequently developed, implemented, and refined Bayesian methods for diagnostic decision making. Ongoing enthusiasm surrounds current work on use of the more general Bayesian belief network approach for clinical diagnosis [1]. Probabilistic systems have played, and will continue to play, an important role in DDSS development.

An additional DDSS alternative exists to categorical (predicate calculus) [39] and probabilistic reasoning that combines features of both but retains a fundamental difference. That alternative is heuristic reasoning, reasoning based on empirical rules of thumb. The HEME program for diagnosis of hematological disorders was one of the earliest systems to employ heuristics and also one of the first systems to use, in effect, criteria tables for diagnosis of disease states. Lipkin, Hardy, Engle, and their colleagues developed HEME in the late 1950s [1, 40–42]. Programs that heuristically match terminology from stored descriptions of disease states to lexical descriptions of patient cases are similar conceptually to HEME. The CONSIDER program developed by Lindberg et al. [43] and the RECONSIDER program developed by Blois and his colleagues [44] used heuristic lexical matching techniques to identify diseases detailed in the Current Medical Information and Terminology (CMIT), a manual of diseases previously compiled and maintained by the American Medical Association. The EXPERT system shell, developed by Weiss and Kulikowski [45], has been used extensively in developing systems that utilize criteria tables, including AI/Rheum [46, 47], for diagnosis of rheumatic disorders, as well as other systems.

G. Anthony Gorry was an enlightened pioneer in the development of heuristic diagnostic systems that employ symbolic reasoning (artificial intelligence, or expert

systems). In a classic paper published in 1968, Gorry [48] outlined the general principles underlying expert system approaches to medical diagnosis that have been incorporated into subsequent systems from the 1970s through the present time. Gorry proposed a formal definition of the diagnostic problem. In a visionary manner, he analyzed the relationships among a generic inference function (used to generate diagnoses from observed findings), a generic test-selection function that dynamically selects the best test to order (in terms of cost and information content), and a pattern-sorting function that is capable of determining if competing diagnoses are members of the same "problem area" (i.e., whether diagnostic hypotheses should be considered together because they are related to pathology in the same organ system). He pointed out the difference between the information value, the economic cost, and the morbidity or mortality risk of performing tests; discussed the cost of misdiagnosis of serious, life-threatening or disabling disorders; noted the potential influence of "red herring" findings on diagnostic systems; described the "multiple diagnosis" problem faced by systems when patients have more than one disease; and suggested that the knowledge bases underlying diagnostic systems could be used to generate simulated cases to test the diagnostic systems.

Gorry's schemata represent the intellectual ancestors of a diverse group of medical diagnostic systems, including, among others: PIP (the Present Illness Program), developed by Pauker et al.; MEDITEL for adult illnesses, which was developed by Waxman and Worley from an earlier pediatric version; INTERNIST-1, developed by Pople, Myers, and Miller; QMR, developed by Miller, Masarie, and Myers; DXplain, developed by Barnett and colleagues; Iliad, developed by Warner and colleagues; the commercial system ISABEL; and a large number of other systems [1, 49–61].

Shortliffe introduced the clinical application of rule-based expert systems for diagnosis and therapy through his development of MYCIN [1, 62] in 1973–1976. MYCIN used backward chaining through its rule base to collect information to identify the organism(s) causing bacteremia or meningitis in patients (see discussion of backward and forward chaining in Chap. 2). A large number of rule-based DDSS have been developed over the years, but most rule-based DDSS have been devoted to narrow application areas due to the extreme complexity of maintaining rule-based systems with more than a few thousand rules [1].

With the advent of the microcomputer came a change in philosophy in regard to the development of DDSS. For example, the global style of diagnostic consultation in the original 1974 INTERNIST-1 program treated the physician-user as unable to solve a diagnostic case [61]. The model assumed that the physician would transfer all historical information, physical examination findings, and laboratory and imaging data to the INTERNIST-1 expert diagnostic consultant program. The physician's subsequent role was that of a passive observer, answering yes or no to questions generated by INTERNIST-1. Ultimately, the omniscient "Greek Oracle" (consultant program) was expected to provide the correct diagnosis and explain its reasoning.

By the late 1980s and early 1990s, DDSS developers abandoned this Greek Oracle mode [9] of diagnostic decision support. For example, the critiquing model developed by Perry Miller [1, 63] and his colleagues, embodied the goal of creating

a combined system that could take advantage of the strengths of both the user's knowledge and the system's abilities.

Several innovative models for computer-assisted medical diagnosis were developed in the 1980s and 1990s. These embodied more formal models that add mathematical rigor to the successful, but more arbitrary, heuristic explorations of the 1970s and early 1980s. However, such models engender tradeoffs, often related to less than perfect underlying data quality, that in many ways make them heuristic as well [64]. Systems based on fuzzy set theory and Bayesian belief networks were developed to overcome limitations of heuristic and simple Bayesian models [1]. Reggia et al. [1, 65] developed set covering models as a formalization of ad hoc problem-area formation (partitioning) schemes, originally described by Gorry in 1968, and later embodied in systems such as Pople's diagnostic algorithms for INTERNIST-1 [66].

Neural networks presented an entirely new approach to medical diagnosis, although the weights learned by simple one-layer networks were analogous or identical to Bayesian probabilities [1]. While neural networks have found applicability in narrow, focused application domains, problems limited their applicability to general diagnosis in broad clinical fields. The difficulties involved selecting the best topology, preventing overtraining and undertraining, and determining what cases to use for training. The more complex a neural network is (number of input and output nodes, number of hidden layers), the greater the need for a large number of appropriate training cases. Often, one cannot obtain large epidemiologically representative data sets that have rigorously determined diagnostic labels. Some developers resort to simulation techniques to generate training cases, but use of artificial cases to train neural networks may lead to suboptimal performance on real cases. Chapters 2 and 3 provide additional detail on the models mentioned above.

11.4 Developing, Implementing, Evaluating, and Maintaining Diagnostic Decision Support Systems

Any successful DDSS must complete a series of developmental stages [2, 67]. First, a new DDSS should meet well-studied and well-documented information needs [4, 5, 68]. Developers must perform a clinical needs assessment to determine the utility of the proposed system and the frequency with which it might be used in various real-world settings. Clinical systems should not be developed simply because a scientist wants to test an exciting new computational algorithm. The rule, "if it's not broke, don't fix it" applies to the development of DDSS, as well as other aspects of technology. Developers must carefully define the scope and nature of the process to be automated. They must also understand the process to be automated well enough to reduce it to an algorithm. All systems, especially DDSS, have boundaries (both in domain coverage and algorithm robustness) beyond which the systems often fail. Developers must understand these limits and make users aware of them—during

DDSS use, if possible. Developers must study DDSS algorithms to determine the ways in which they might fail, both due to inherent limitations and to flaws that might occur during the processes of implementation and use [2].

Evaluation must first occur carefully, initially "in vitro" (outside of the patient care arena, with no risks to patients), and, once warranted, in vivo (prospectively, on the front lines of actual patient care delivery) in order to determine if the DDSS improves or promotes important outcomes that are not possible with the pre-existing manual system [69]. Finally, developers and users must demonstrate the practical utility of the system by showing that clinicians can adopt it for productive daily use [2]. A potentially great system that is not used cannot have a beneficial impact on clinical outcomes. Unfortunately, few existing DDSS have yet fulfilled these criteria.

A number of problems have limited the ultimate success of DDSS to date. These include: difficulties with domain selection and knowledge base construction and maintenance; problems with the diagnostic algorithms and user interfaces; the problem of system evolution, including evaluation, testing, and quality control; issues related to machine interfaces and clinical vocabularies; and legal and ethical issues. These issues are discussed below.

11.4.1 Clinical Domain Selection

DDSS domain selection can pose problems. Substantial clinical domains require construction of corresponding, high-quality DDSS knowledge bases. Their construction and maintenance can consume dozens of person-years of effort in broad domains such as general internal medicine. To date, most large DDSS knowledge bases have at least initially been created in the academic environment. Many projects do not have adequate funding to sustain such activity over time [70]. Availability of adequate domain expertise is also a problem. Clinical collaborators generally earn their wages through patient care or clinical research, and sustaining high-level input from individuals with adequate clinical expertise can be difficult in the face of real-world demands. Commercial vendors must hire an adequate and well qualified staff of physicians in order to maintain medical knowledge bases. The number of users willing to purchase a DDSS program and its updates, as well as the price they are willing to pay, limit the income generated through the sale of the DDSS. Obtaining a critical volume of sales to support ongoing developments and updates is difficult.

Different types of problems afflict DDSS that target narrow domains. One problem is garnering an adequate audience. The CASNET system was an exemplary prototypic system for reasoning pathophysiologically about the diagnosis and therapy of glaucoma [71]. It typifies a problem that can occur with successful focal experimental expert systems with limited scope—the persons most likely to require such a specialized system's use in clinical medicine are the domain experts whose knowledge was used to develop the system. The persons who routinely diagnose and treat glaucoma are ophthalmologists, who are by definition board-certified

specialists in the domain of ophthalmology. Programs like CASNET, in effect, run the risk of preaching to the choir. It is more difficult for an automated system to provide useful expertise in a given narrow specialty; human subspecialists in that area may rightly or wrongly believe they need not use it. Conversely, generalists are also unlikely to use a system with very narrow range of function. Specialty-specific, focused DDSS programs like the CASNET system must be extremely robust and provide more than one kind of service (e.g., by providing integrated record management and other functions in addition to DDSS functionality) in order to find use in clinical practice.

11.4.2 Knowledge Base Construction and Maintenance

Knowledge base maintenance is critical to the clinical validity of a DDSS [1]. Yet it is hard to determine when new clinical information becomes established as "fact." First reports of new clinical discoveries in highly regarded medical journals must await confirmation by other groups over time before their content can be added to a medical knowledge base. The nosological labels used in diagnosis reflect the current level of scientific understanding of pathophysiology and disease. They may change over time without the patient or the patient's illness, per se, changing [1]. For example, changes occur in how a label is applied when the "gold standard" for making a diagnosis shifts from a pathological biopsy result to an abnormal serological or genetic test—patients with earlier, previously unrecognized forms of the illness may be labeled as having the disease. Corresponding changes must be made to keep a DDSS knowledge base up to date.

Knowledge base construction must become a scientifically reproducible process that qualified individuals can successfully undertake at any site [72]. Knowledge base construction should be clinically grounded, based on objective, peer-reviewed information (e.g., literature-based) whenever possible. Attempts to "tune" a DDSS knowledge base to improve DDSS performance on a given case or group of cases should be strongly discouraged. A general system tuned in that manner lacks lasting calibration across all cases—changes improving performance for one specific case may degrade performance on other previously diagnosable cases. Any updates should have an objective basis, such as information culled from the medical literature.

If the process of knowledge base construction is highly dependent on a single individual, or can only be carried out at a single institution, then the survival of that system over time is in jeopardy. While much of the glamour of computer- based diagnostic systems lies in the computer algorithms and interfaces, the long-term value and viability of a system depends on the quality, accuracy, and timeliness of its knowledge base [1].

Even initially successful DDSS cannot survive unless the medical knowledge bases supporting them are kept current. This can require Herculean efforts. Shortliffe's MYCIN program [62] was developed as a research project to demon-

strate the applicability of rule-based expert systems to clinical medicine. MYCIN was a brilliant, pioneering effort in this regard. The evaluation of MYCIN in the late 1970s by Yu and colleagues demonstrated that the program could perform at the expert level on challenging cases [73]. But MYCIN was never put into routine clinical use, nor was an effort made to update its knowledge base over time. After 1980, lack of maintenance led its antibiotic therapy knowledge base to become out of date.

11.4.3 Diagnostic Algorithms and User Interfaces

Just as computer-based implementation of many complex algorithms involves making trade-offs between space (memory) and time (CPU cycles), development of real-world diagnostic systems involves a constant balancing of theory (model complexity) and practicality (ability to construct and maintain adequate medical databases or knowledge bases, and ability to create systems which respond to users' needs in an acceptably short time interval) [64]. We may understand, in theory, how to develop systems that take into account gradations of symptoms, the degree of uncertainty in the patient and/or physician-user regarding a finding, the severity of each illness under consideration, the pathophysiological mechanisms of disease, and/or the time course of illnesses. Such complexities may ultimately be required to make actual systems work reliably. Nevertheless, it is not yet practical to build such complex, broad-based systems for patient care. The effort required to build and maintain superficial knowledge bases is measured in dozens of person-years of effort, and more complex knowledge bases are likely to require an order of magnitude greater effort [1]. The evidence to support many fine-grained diagnostic knowledge representation schemes may not yet exist in objective repositories such as the peer-reviewed literature.

Although some have posited that DDSS will eventually replace physicians as primary diagnosticians [74], that position seems untenable. A clinician cannot easily convey his or her complete understanding of a complex patient case to a computer program. One should never assume that a computer program "knows" all that needs to be known about a patient case, no matter how much time and effort is spent on data input. As a result, the clinician-user who directly evaluated the patient must be considered to be the definitive source of information about the patient during the entire course of any computer-based consultation [2]. In addition, the highly skilled health care practitioner—who understands the patient as a person—possesses the most important intellect to be employed during a consultation. That user should control the intellectual process of computer-based consultation, determining the sequence of steps to take place, which questions to pose, and whether those questions have been addressed. Systems must provide flexible environments that adapt to the user's needs and problems, rather than providing an interface that is inflexible and which penalizes the user for deviating from the normal order of system operation.

11.4.4 Testing, Evaluation, and Quality Control

System evaluation in biomedical informatics should take place as an ongoing, strategically planned process, not as a single event or small number of episodes [67, 69]. Complex software systems and accepted medical practices both evolve rapidly, so evaluators and readers of evaluations face moving targets. As previously noted, systems are of value only when they help users to solve users' problems. Users, not systems, characterize and solve clinical diagnostic problems. In keeping with that observation—that the DDSS user defines the problem to be solved—the ultimate unit of evaluation should be whether the user plus the system is better than the unaided user with respect to a specified task or problem (usually one generated by the user) [2, 69, 75].

Extremely important during system development are lessons learned (and modifications) based on informal formative evaluations. Developers of DDSS should analyze new DDSS cases on a regular (e.g., weekly) basis. After each failure of the DDSS to make a "correct" diagnosis, careful analysis of both the system's knowledge base and diagnostic algorithms must be carried out. Both the information in the knowledge base on the "correct" diagnosis, and the information on any diagnoses offered in error, must be reviewed and potentially updated. Updates should be evidence-based, not just arbitrary "tuning" of the system for a specific problematic case. In addition, periodic rerunning of all previous test cases, done on an annual (or similar) basis, can verify that no significant "drift" in either the knowledge base or the diagnostic programs have occurred.

Formal evaluations of DDSS should take into account the following four perspectives: (1) appropriate evaluation design; (2) specification of criteria for determining DDSS efficacy in the evaluation; (3) evaluation of the boundaries or limitations of the DDSS; and (4) identification of potential reasons for "lack of system effect" [69]. Each of these issues is discussed below.

Appropriate Evaluation Design

Evaluation plans should be appropriate for the information needs being addressed, the level of system maturity, and users' intended form of DDSS usage (or specific system function evaluated) [67, 69]. The same DDSS may serve as an electronic textbook for one user, a diagnostic checklist generator for another user, a consultant to determine the next useful step in a specific patient's evaluation for a third user, and a tool to critique/reinforce the users' own pre-existing hypotheses for a fourth user. Each system function would require a different form of evaluation whenever anticipated user benefits depend on which system function is used. Evaluations should clearly state which user objective is being studied and which of the available system functions are relevant to that objective.

In 1994, Berner and colleagues evaluated the ability of several systems to generate first-pass differential diagnoses from a fixed set of input findings [76]. These findings were not generated by everyday clinical users, but from written case summaries of real patient data. That approach was dictated by the desire to standardize system inputs and outputs for purposes of multisystem use. The primary goal of Berner et al. was to develop methods and metrics that would characterize aspects of system performance in a manner useful for rationally comparing different systems and their functions. All of the systems in that study were capable of generating questions to further refine the initial differential diagnoses, which is the intended mode of clinical use for such systems. Because that study was not intended to produce a definitive rating or comparison of the systems themselves, the involved systems were not placed in the hands of end users, nor were the systems used in a manner to address common end-user needs. Even though the evaluation did not examine this capability, the methods used by Berner were sound. Generating a first-pass differential diagnosis is a good initial step, but subsequent evidence gathering, reflection, and refinement are required.

There are important questions that must be answered in the evaluation. Are the problems ones that clinical users generate during clinical practice, or artificial problems generated by the study design team? Is the case material accurately based on actual patient cases? Note that there can be no truly verifiable diagnosis when artificial, manually constructed or computer-generated cases are used. Are the evaluation subjects clinical users whose participation occurs in the clinical context of caring for the patients used as "test cases?" Are clinical users evaluating abstracts of cases they have never seen, or are nonclinical personnel evaluating abstracted clinical cases using computer systems? Are users free to use all system components in whatever manner they choose, or is it likely that the study design will constrain users to exercise only limited components of the system? The answers to these questions will determine the generalizability of the results of the evaluation.

Specification of Criteria for Determining Efficacy in the Evaluation

Evaluations must identify criteria for "successful" system performance similar to what clinical practitioners would use during actual practice. Diagnosis, or more properly, "diagnostic benefit," must be defined in such contexts. Similarly, what it means to establish a diagnosis must be carefully defined. For example, it is not adequate to accept hospital discharge diagnoses at face value as a "gold standard" since discharge diagnoses are not of uniform quality—they have been documented to be influenced by physician competency, coding errors, and economic pressures. Furthermore, some discharge diagnoses may be "active" (undiagnosed at admission and related to the patient's reason for hospitalization), while others may be relevant but inactive. Criteria for the establishment of a "gold standard" diagnosis should be stated prospectively, before beginning data collection.

Evaluation of the Boundaries or Limitations

A system may fail when presented with cases outside its knowledge base domain, but if an evaluation uses only cases from within that domain, this failure may never be identified. The limits of a system's knowledge base are a concern because patients do not accurately triage themselves to present to the most appropriate specialists. For instance, as discussed earlier, de Dombal's abdominal pain system performed very well when used by surgeons to determine if patients presenting with abdominal pain required surgery [10]. However, a patient with atypical appendicitis may present to an internist, and a patient with abdominal pain due to lead poisoning may first see a surgeon.

Identification of Potential Reasons for "Lack of System Effect"

DDSS operate within a system that not only includes the DDSS itself, but also the user and the healthcare environment in which the user practices. A model of all of the possible influences on the evaluation outcomes would include DDSS-related factors (knowledge base inadequacies, inadequate synonyms within vocabularies, faulty algorithms, etc.), user-related factors (lack of training or experience with the system, failure to use or understand certain system functions, lack of medical knowledge or clinical expertise, etc.) and external variables (lack of available gold standards, failure of patients or clinicians to follow-up during study period). It is important to recognize that studies that focus on one aspect of system function may have to make compromises with respect to other system or user-related factors in order to have an interpretable result. Additionally, in any DDSS evaluation, the user's ability to generate meaningful input into the system, and the system's ability to respond to variable quality of input from different users, is an important concern.

Evaluations of DDSS must each take a standard objective (which may be only one component of system function) and measure how effectively the system enhances users' performances, using a study design that incorporates the most appropriate and rigorous methodology relative to the stage of system development. The ultimate clinical end user of a given DDSS must determine if published evaluation studies examine the system's function in the manner that the user intends to use it. This is analogous to a practitioner determining if a given clinical trial (of an intervention) is relevant to a specific patient by matching the given patient's characteristics to the study's inclusion and exclusion criteria, population demographics, and the patient's tolerance for the proposed forms of therapy as compared to alternatives. The reporting of an individual "negative study" of system performance should not, as it often does now, carry the implication that the system is globally suboptimal. A negative result for one system function does not mean that, for the same system, some users cannot derive significant benefits for other system functions. Similarly, complete evaluation of a system over time should examine basic components (e.g., the knowledge base, ability to generate reasonable differential diagno-

ses, ability to critique diagnoses, and so on), as well as clinical functionality (e.g., can novice users, after standard training, successfully employ the system to solve problems that they might not otherwise solve as efficiently or completely?). The field of DDSS evaluation will become mature only when clinical system users regularly derive the same benefit from published DDSS evaluations as they do from evaluations of standard clinical interventions.

11.4.5 Interface and Vocabulary Issues

A critical issue for the success of large-scale, generic DDSS is their environment. Small, limited, "niche" systems may be adopted and used by the focused community for which they are intended, while physicians in general medical practice, for whom the large-scale systems are intended, may not perceive the need for diagnostic assistance on a frequent enough basis to justify purchase of one or more such systems. Therefore, it is common wisdom that DDSS are most likely to succeed if they can be integrated into a clinical environment so that patient data capture is already performed by automated laboratory and/or hospital information systems. In such an environment, the physician will not have to manually enter all of a patient's data in order to obtain a diagnostic consultation. However, automated transfer of all the information about a patient from a hospital information system to a diagnostic consultation system is nontrivial. If 100 hematocrits were measured during a patient's admission, which one(s) should be transferred to the consultation system—the mean, the extremes, or the value typical for a given time in a patient's illness? Should all findings be transferred to the consultation system, or only those findings relevant to the patient's current illness? These questions must be resolved by careful study before one can expect to obtain patient consultations routinely and automatically within the context of a hospital information system. Another reason for providing an integrated environment is that users will not use a system unless it is sufficiently convenient to do so. By integrating DDSS into healthcare provider results reporting and order entry systems, the usual computer-free workflow processes of the clinician can be replaced with an environment conducive to accomplishing a number of computer-assisted clinical tasks, making it more likely that a DDSS will be used.

Interfaces between automated systems are, at times, as important as the man-machine interface [77, 78]. Fundamental questions, such as the definition of diseases and of findings, limit our ability to combine data from the literature, from clinical databanks, from hospital information systems, and from individual experts' experiences in order to create DDSS. Similar problems exist when trying to match the records from a given case with a computer-based diagnostic system. A diagnostic system may embody different definitions for patient descriptors than those of the physician who evaluated the patient, even though the words used by each may be identical.

In order to facilitate data exchange among local and remote programs, it is mandatory to have a lexicon or interlingua which facilitates accurate and reliable transfer of information among systems that have different internal vocabularies (data dictionaries). The United States National Library of Medicine Unified Medical Language System (UMLS) project, which started in 1987 and continues through the present time, represents one such effort [79].

11.4.6 Legal and Ethical Issues

Proposals have suggested that governmental agencies, such as the United States Food and Drug Administration (FDA), which oversees medical devices, regulate use of clinical software programs such as DDSS. These proposals include a variety of recommendations that manufacturers of such systems would be required to perform to guarantee that the systems would function per specifications.

There is debate about whether these consultation systems are actually devices in the same sense as other regulatable devices. In the past, governmental regulation has not been considered necessary when a licensed practitioner is the user of a DDSS [80]. It would be both costly and difficult for the government to regulate DDSS more directly, even if a decision were made to do so. For general DDSS programs like Iliad, QMR, Meditel and DXplain, with hundreds to thousands of possible diagnoses represented in their knowledge bases [76], conducting prospective clinical trials, to demonstrate that the system worked for all ranges of diagnostic difficulty for a variety of patients with each diagnosis, would require enrollment of huge numbers of patients and would cost millions of dollars.

Other approaches, such as a "software quality audit" to determine, prospectively, if a given software product has flaws would also be clinically impractical. The clinician seeking help may have any of several dozen kinds of diagnostic problems in any given case. Unless it is known, for a given case, which kind of problem the practitioner will have, performing a software quality audit could not predict if the system would be useful.

Consider the dilemma the FDA or other responsible regulatory agency would face if it agreed to review situations when a user files a complaint. First, one must note that few patients undergo definitive enough diagnostic evaluations to make it possible to have a "gold standard" (certain) diagnosis. So if the doctor claims the program was wrong, a major question would be how governmental auditors would know what the actual "right" diagnosis was. Second, the reviewers would need to know all of the information that was knowable about the patient at the time the disputed diagnosis was offered. This could potentially violate patient confidentiality if the records were sent to outsiders for review. All sources of information about the patient would have to be audited, and this could become as difficult as evidence gathering in a malpractice trial. To complete the sort of audit described, the governmental agency would have to determine if the user had been appropriately trained and if the user used the program correctly. Unless the program had an internally

stored complete audit trail of each session (down to the level of saving each keystroke the user typed), the auditors might never be able to recreate the session in question. Also, the auditors would have to study whether the program's knowledge base was appropriate. Initial development of the R-QMR knowledge base at the University of Pittsburgh required an average of three person-weeks of a clinician's time, which went into literature review of 50–150 primary articles about each disease, with additional time for synthesis and testing against cases of real patients with the disease. For an auditor to hire the required expertise to review this process for hundreds to thousands of diseases for each of the programs that it would have to review and subsequently monitor would be costly and cumbersome. The ultimate question, very difficult to answer, would be whether the original user in the case in question used the system in the best way possible for the given case. Making such a determination would require the governmental agency to become expert in the use of each DDSS program. This could take up to several months of training and practice for a single auditor to become facile in the use of a single system. It would be difficult for a governmental agency to muster the necessary resources for even a small number of such complaints, let alone nationwide for multiple products with thousands of users. The complexity of these issues makes it very difficult to formulate appropriate regulatory policy. In addition to legal issues concerning regulation, there are other legal and ethical issues relating to use of DDSS that are discussed in Chap. 8.

11.5 Diagnostic Decision Support Systems Circa 2015

Recent emphasis on preventable errors in clinical practice originated in the 1980s with published studies on adverse drug effects, and peaked with the Institute of Medicine's more comprehensive report, To Err Is Human [81]. Many researchers neglected or downplayed the frequency and importance of diagnostic errors, especially in the outpatient setting, because little was known at the time. Recently, increased interest has focused on diagnostic errors and their prevention [82–92]. The Society to Improve Diagnosis In Medicine (SIDM) grew out of the momentum generated by post-2000 annual conferences on diagnostic errors. In 2014, SIDM began publishing a journal, Diagnosis [92]. In 2015, the Agency for Healthcare Research and Quality (AHRQ) emphasized the importance of diagnosis by issuing new RFAs for methods to reduce diagnostic errors in the outpatient setting. The Institute of Medicine (National Academy of Medicine) of the National Academy of Sciences published its summary of a multi-year study of diagnostic errors [93]. The potential for implementation of DDSS in clinical practice, and the ability to study their impact has never been greater.

Three general, non-focal DDSS available in 2015 merit mention as exemplars: VisualDx® [94–99], DXplain [54, 56, 57, 100], and ISABEL [60, 101–105]. VisualDx® and ISABEL are marketed commercially; DXplain is available via institutional licenses for an annual fee. The web-based DXplain DDSS represents the

current evolution of a system initially developed in 1984 by G. Octo Barnett and colleagues at the Massachusetts General Hospital [100]. Dr. Barnett, the primary developer of DXplain, often stated that the inspiration for the system grew out of his respect for INTERNIST-1 [58, 59]. According to the 2015 DXplain web site, "the current DXplain knowledge base (KB) includes over 2400 diseases and over 5000 clinical findings (symptoms, signs, epidemiologic data and laboratory, endoscopic and radiologic findings)" [100]. The ISABEL DDSS was developed as a commercial application from the outset. It originally covered Pediatric diagnosis [101–104] and its knowledge base has grown to now include adult disorders. In 2003, the developers of ISABEL published an evaluation of ISABEL, proposing that previous rigorous standards for DDSS evaluation might be unnecessary [105]. Berner discussed the implications of evaluating DDSS using less than absolute gold standards, as was proposed by the ISABEL team, in a well-balanced perspective covering "correctness" of diagnosis, "appropriateness" of management suggestions, end-user acceptance and satisfaction, degree of adoption and use of a DDSS, and issues related to human-computer system interfaces [106]. Like many heuristic systems before them, DXplain and ISABEL behaviorally follow Gorry's 1968 DDSS template.

A current DDSS that satisfies many of the previously discussed desiderata for a creating, maintaining, and distributing a successful system is VisualDx, developed by Dr. Art Papier and colleagues [94–99]. Dr. Papier is an academically-based dermatologist who has developed an extensive consortium of collaborating institutions to construct and maintain the VisualDx knowledge base, consisting of dermatological images, a standardized lexicon of text descriptions for each of the images, and summary characterizations of the disorders associated with each image and with each text description. The web site "visualdx.com" [94] states the following: "VisualDx is a diagnostic clinical decision support and reference tool that combines high-quality, peer-reviewed medical images and expert information to support today's internists and infectious disease physicians in the accurate recognition and management of disease …over 1500 hospitals and large clinics … recognize VisualDx as a … quality and safety system."

11.6 The Future of Diagnostic Decision Support Systems

It is relatively safe to predict that specialized, focused DDSS will proliferate, and a sizable number of them will find widespread application [1]. As new medical devices are developed and older devices automated, DDSS software that enhances the performance of the device, or helps users to interpret the output of the device, will become essential.

Computerized electrocardiogram (ECG) analysis, automated arterial blood gas interpretation, automated protein electrophoresis reports, and automated differential blood cell counters, are but a few examples of such success at the present time. Since Miller's 1994 article summarizing past DDSS developmental activities [1], the great majority of new articles on "diagnosis, computer-assisted" indexed in

MEDLINE have described focused systems for the interpretation of images (radiological studies and pathology cytology/sections/slides), signals (ECGs, electroencephalograms (EEGs), and so on), and diagnosis of very narrowly defined clinical conditions. One by-product of the success of these systems is that users may be less vigilant in questioning system accuracy. In a 2003 article, Tsai and colleagues pointed out the potential clinical dangers of overreliance of inexpert clinicians on computer systems for advice—they tend to follow the advice even when it is wrong [107].

For the foreseeable future, machine-learning approaches to DDSS will find success in the realm of specialized, focused systems. There, adequate training exemplars can be found, and the number of categories to discriminate is relatively small (typically dozens). A somewhat related example, IBM's Watson™ analytic engine [108], is also more likely to find success in DDSS applications in focal domains rather than general diagnosis for medicine or pediatrics. Watson uses natural language processing to draw statistical relationships among terms extracted from textual documents [108], such as the biomedical literature or patients' charts. In reviewing the literature for the INTERNIST-1 and QMR projects, project members noticed that human expertise at a high level must resolve among conflicting reports as to whether a given disorder causes a given finding. Furthermore, in a given case summary appearing in either the literature or an EMR, patients often have multiple conditions that can cause or, in combination, exacerbate a given finding—e.g., a patient with shortness of breath who has both emphysema and heart failure. Purely automated systems would likely experience more difficulty than an expert clinician in sorting out which disorder caused the finding on an algorithmic basis. Furthermore, for extremely rare disorders, such as primary sarcoma of the heart, a sufficient number of case reports may not exist for algorithmic extraction of findings with certainty. The whole field of meta-analysis, which attempts to determine from published randomized controlled trials the quality of evidence supporting various therapeutic approaches to a given disorder, indicates the complexity of decision-making involved in collating evidence. Machine learning and Watson-like attempts to summarize the literature on diagnosis, which lacks the rigor of randomized controlled trials, will also encounter extreme difficulty when attempting to derive evidence bases to support DDSS in broad fields such as medicine or pediatrics.

So manual, or quasi-manual approaches to DDSS knowledge base curation by qualified clinical experts will remain the best method to construct and maintain DDSS knowledge bases in the near-term future. Watson-like systems may, however, provide useful assistance to humans or heuristic DDSS in general clinical domains by, upon request, searching for evidence supporting (or refuting) a given specific diagnosis within a single patient's voluminous EMR record.

The future of large-scale, "generic" diagnostic systems is hopeful, although less certain. As discussed in this and other chapters, a small number of large-scale, generic DDSS are in limited use in clinical practice. Systems like VisualDx provide hope that a model for ongoing maintenance and distribution for DDSS can be feasible. Nevertheless, it is well established that DDSS can play a valuable role in medical education [1]. The process of knowledge base construction, utilization of

such knowledge bases for medical education in the form of patient case simulations, and the use of DDSS have all been shown to be of educational value in a variety of institutional settings.

In summary, the future of DDSS appears to be promising. The number of researchers in the field is growing. The diversity of DDSS is increasing. The number of commercial enterprises interested in DDSS is expanding. Rapid improvements in computer technology continue to be made. A growing demand for cost-effective clinical information management, and the desire for better health care, is sweeping the United States [109]. Evidence-based medicine is now in vogue. All these factors will insure that new and productive DDSS applications will be developed, evaluated, and used.

References

1. Miller RA. Medical diagnostic decision support systems—past, present, and future: a threaded bibliography and commentary. J Am Med Inform Assoc. 1994;1:8–27.
2. Miller RA. Why the standard view is standard: people, not machines, understand patients' problems. J Med Philos. 1990;15:581–91.
3. Flexner SB, Stein J, editors. The Random House college dictionary. Revised ed. New York: Random House; 1988. p. 366.
4. Osheroff JA, Forsythe DE, Buchanan BG, Bankowitz RA, Blumenfeld BH, Miller RA. Physicians' information needs: an analysis of questions posed during clinical teaching in internal medicine. Ann Intern Med. 1991;114:576–81.
5. Forsythe DE, Buchanan BG, Osheroff JA, Miller RA. Expanding the concept of medical information: an observational study of physicians' information needs. Comput Biomed Res. 1992;25:181–200.
6. Miller R, Masarie FE, Myers J. Quick Medical Reference (QMR) for diagnostic assistance. MD Comput. 1986;3:34–8.
7. Miller RA, Masarie FE Jr. The quick medical reference (QMR) relationships function: description and evaluation of a simple, efficient "multiple diagnoses" algorithm. Medinfo. 1992;512–18.
8. Pople Jr HE. Heuristic methods for imposing structure on ill-structured problems: the structuring of medical diagnostics. In: Szolovits P, editor. Artificial intelligence in medicine, AAAS symposium series. Boulder: Westview Press; 1982. p. 119–90.
9. Miller RA, Masarie Jr FE. The demise of the "Greek Oracle" model for medical diagnosis systems. Methods Inf Med. 1990;29:1–2.
10. de Dombal FT, Leaper DJ, Horrocks JC, Staniland JR, McCann AP. Human and computer-aided diagnosis of abdominal pain: further report with emphasis on performance of clinicians. Br Med J. 1974;1:376–80.
11. Adams ID, Chan M, Clifford PC, et al. Computer aided diagnosis of acute abdominal pain: a multicentre study. Br Med J (Clin Res Ed). 1986;293:800–4.
12. Newell A, Simon HA. Human problem solving. Englewood Cliffs: Prentice Hall; 1972.
13. Kahneman D, Slovic P, Tversky A, editors. Judgment under uncertainty: heuristics and biases. Cambridge, UK: Cambridge University Press; 1982.
14. Elstein AS, Shulman LS, Sprafka SA. Medical problem solving: an analysis of clinical reasoning. Cambridge, MA: Harvard University Press; 1978.
15. Kassirer JP, Gorry GA. Clinical problem-solving—a behavioral analysis. Ann Intern Med. 1978;89:245–55.

16. Miller GA. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol Rev. 1956;63:81–97.
17. Evans DA, Patel VL, editors. Cognitive science in medicine. Cambridge, MA: MIT Press; 1989.
18. Simon HA. The structure of ill-structured problems. Artif Intell. 1973;4:181–201.
19. Miller RA, Pople Jr HE, Myers J. INTERNIST-1, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med. 1982;307:468–76.
20. Eddy DM, Clanton CH. The art of diagnosis: solving the clinicopathological conference. N Engl J Med. 1982;306:1263–9.
21. Weed LL. Medical records that guide and teach. N Engl J Med. 1968;278:593–600. 652–657.
22. Hurst JW, Walker HK, editors. The problem-oriented system. New York: Medcom Learning Systems; 1972.
23. Kassirer JP. Diagnostic reasoning. Ann Intern Med. 1989;110:893–900.
24. Brunswik E. Representative design and probabilistic theory in a functional psychology. Psychol Rev. 1955;62:193–217.
25. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason. Science. 1959;130:9–21.
26. Warner HR, Toronto AF, Veasey LG, Stephenson R. Mathematical approach to medical diagnosis. JAMA. 1961;177:75–81.
27. Raiffa H. Decision analysis. Reading: Addison-Wesley; 1970.
28. Pauker SG, Kassirer JP. Decision analysis. N Engl J Med. 1987;316:250–8.
29. Dawes RM, Faust D, Meehl PE. Clinical versus actuarial judgment. Science. 1989;243:1668–74.
30. Gigerenzer G, Hoffrage U. How to improve Bayesian reasoning without instruction: frequency formats. Psychol Rev. 1995;102:684–704.
31. Feinstein AR. An analysis of diagnostic reasoning. I. The domains and disorders of clinical macrobiology. Yale J Biol Med. 1973;46:212–32.
32. Feinstein AR. An analysis of diagnostic reasoning. II. The strategy of intermediate decisions. Yale J Biol Med. 1973;46:264–83.
33. Horvitz EJ, Heckerman DE. The inconsistent use of measures of certainty in artificial intelligence research. In: Kanal LN, Lemmer JF, editors. Uncertainty in artificial intelligence, vol. 1. Amsterdam: Elsevier Science; 1986. p. 137–51.
34. Blois MS. Clinical judgment and computers. N Engl J Med. 1980;303:192–7.
35. Shortliffe EH, Buchanan BG, Feigenbaum EA. Knowledge engineering for medical decision-making: a review of computer-based clinical decision aids. Proc IEEE. 1979;67:1207–24.
36. Willems JL, Abreu-Lima C, Arnaud P, et al. The diagnostic performance of computer programs for the interpretation of electrocardiograms. N Engl J Med. 1991;325:1767–73.
37. Bleich HL. Computer evaluation of acid-base disorders. J Clin Invest. 1969;48:1689–96.
38. Gorry GA, Barnett GO. Experience with a model of sequential diagnosis. Comput Biomed Res. 1968;1:490–507.
39. Szolovits P, Pauker SG. Categorical and probabilistic reasoning in medical diagnosis. Artif Intell. 1978;11:114–44.
40. Lipkin M, Hardy JD. Differential diagnosis of hematological diseases aided by mechanical correlation of data. Science. 1957;125:551–2.
41. Lipkin M, Hardy JD. Mechanical correlation of data in differential diagnosis of hematological diseases. JAMA. 1958;166:113–23.
42. Lipkin M, Engle Jr RL, Davis BJ, Zworykin VK, Ebald R, Sendrow M. Digital computer as aid to differential diagnosis. Arch Intern Med. 1961;108:56–72.
43. Lindberg DAB, Rowland LR, Buch CR Jr, Morse WF, Morse SS. Consider: a computer program for medical instruction. In: Proceedings of the ninth IBM medical symposium. White

Plains: IBM, 1968. Conference dates: 9th IBM Medical Symposium, Burlington, Vermont, October 24–26, 1969.
44. Nelson SJ, Blois MS, Tuttle MS, et al. Evaluating RECONSIDER: a computer program for diagnostic prompting. J Med Syst. 1985;9:379–88.
45. Weiss S, Kulikowski CA. EXPERT: a system for developing consultation models. In: Proceedings of the sixth international joint conference on artificial intelligence. Tokyo; 1979.
46. Lindberg DAB, Sharp GC, Kingsland III LC, et al. Computer-based rheumatology consultant. In: Linberg DAB, Kaihara S, editors. Proceedings of MEDINFO 80 Tokyo, third world conference on medical informatics. Amsterdam: North Holland Publishing Company; 1980. p. 1311–5.
47. Moens HJ, van der Korst JK. Development and validation of a computer program using Bayes' theorem to support diagnosis of rheumatic disorders. Ann Rheum Dis. 1992;51:266–71.
48. Gorry A. Strategies for computer-aided diagnosis. Math Biosci. 1968;2:293–318.
49. Pauker SG, Gorry GA, Kassirer JP, Schwartz WB. Towards the simulation of clinical cognition. Taking a present illness by computer. Am J Med. 1976;60:981–96.
50. Waxman HS, Worley WE. Computer-assisted adult medical diagnosis: subject review and evaluation of a new microcomputer-based system. Medicine. 1990;69:125–36.
51. Pople HE, Myers JD, Miller RA. DIALOG: a model of diagnostic logic for internal medicine. In: Proceedings of the fourth international joint conference on artificial intelligence. Tiblisi; 1975. p. 848–55.
52. First MB, Soffer LJ, Miller RA. QUICK (Quick Index to Caduceus Knowledge): using the INTERNIST-1/Caduceus knowledge base as an electronic textbook of medicine. Comput Biomed Res. 1985;18:137–65.
53. Miller RA, McNeil MA, Challinor S, Masarie Jr FE, Myers JD. Status report: the Internist-1/Quick Medical Reference project. West J Med. 1986;145:816–22.
54. Hupp JA, Cimino JJ, Hoffer EF, Lowe HJ, Barnett GO. DXplain—a computer- based diagnostic knowledge base. In: Proceedings of the fifth world conference on medical informatics (MEDINFO 86). Amsterdam. p. 117–21.
55. Warner HR, Haug P, Bouhaddou O, Lincoln M. ILIAD as an expert consultant to teach differential diagnosis. In: Greenes RA, editor. Proceedings of the twelfth annual symposium on computer applications in medical care. Los Angeles: IEEE Computer Society; 1988. p. 371–6.
56. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain. An evolving diagnostic decision-support system. JAMA. 1987;258:67–74.
57. Elkin PL, Liebow M, Bauer BA, Chaliki S, Wahner-Roedler D, Bundrick J, Lee M, Brown SH, Froehling D, Bailey K, Famiglietti K, Kim R, Hoffer E, Feldman M, Barnett GO. The introduction of a diagnostic decision support system (DXplain™) into the workflow of a teaching hospital service can decrease the cost of service for diagnostically challenging Diagnostic Related Groups (DRGs). Int J Med Inform. 2010;79(11):772–7.
58. Barnett GO. The computer and clinical judgment. N Engl J Med. 1982;307(8):493–4.
59. Barnett GO. Personal communications to RA Miller, 1988–2002.
60. Graber ML, Mathew A. Performance of a web-based clinical diagnosis support system for internists. J Gen Intern Med. 2008;23 Suppl 1:37–40.
61. Miller RA. A history of the INTERNIST-1 and Quick Medical Reference (QMR) computer-assisted diagnosis projects, with lessons learned. Yearb Med Inform. 2010;121–136.
62. Shortliffe EH. Computer-based medical consultations: MYCIN. New York: Elsevier; 1976.
63. Miller PL. A critiquing approach to expert computer advice: ATTENDING. Boston: Pittman; 1984.
64. Aliferis CF, Miller RA. On the heuristic nature of medical decision support systems. Methods Inf Med. 1995;34:5–14.
65. Reggia JA, Nau DS, Wang PY. Diagnostic expert systems based on a set covering model. Int J Man Mach Stud. 1983;19:437–60.

66. Berman L, Miller RA. Problem area formation as an element of computer aided diagnosis: a comparison of two strategies within quick medical reference (QMR). Methods Inf Med. 1991;30:90–5.
67. Stead WW, Haynes RB, Fuller S, et al. Designing medical informatics research and library-resource projects to increase what is learned. J Am Med Inform Assoc. 1994;1:28–33.
68. Covell DG, Uman GC, Manning PR. Information needs in office practice: are they being met? Ann Intern Med. 1985;103:596–9.
69. Miller RA. Evaluating evaluations of medical diagnostic systems. J Am Med Inform Assoc. 1996;3:429–31.
70. Yu VL. Conceptual obstacles in computerized medical diagnosis. J Med Philos. 1983;8:67–75.
71. Weiss S, Kulikowski CA, Safir A. Glaucoma consultation by computer. Comput Biol Med. 1978;8:24–40.
72. Giuse NB, Giuse DA, Miller RA, et al. Evaluating consensus among physicians in medical knowledge base construction. Methods Inf Med. 1993;32:137–45.
73. Yu VL, Fagan LM, Wraith SM, et al. Antimicrobial selection by computer: a blinded evaluation by infectious disease experts. JAMA. 1979;242:1279–82.
74. Mazoue JG. Diagnosis without doctors. J Med Philos. 1990;15:559–79.
75. Friedman CP. A "fundamental theorem" of biomedical informatics. J Am Med Inform Assoc. 2009;16(2):169–70.
76. Berner ES, Webster GD, Shugerman AA, et al. Performance of four computer-based diagnostic systems. N Engl J Med. 1994;330:1792–6.
77. Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. J Am Med Inform Assoc. 2006;13(3):277–88.
78. Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. A model for evaluating interface terminologies. J Am Med Inform Assoc. 2008;15(1):65–76.
79. Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. Methods Inf Med. 1993;32:281–91.
80. Young FE. Validation of medical software: present policy of the Food and Drug Administration. Ann Intern Med. 1987;106:628–9.
81. Kohn LT, Corrigan JM, Donaldson MS, editors. To err is human: building a safer health system. Washington, DC: National Academy Press; 2000.
82. Friedman CP, Gatti GG, Franz TM, et al. Do physicians know when their diagnoses are correct? Implications for decision support and error reduction. J Gen Intern Med. 2005;20:334–9.
83. Shojania KG, Burton EC, McDonald KM, et al. Changes in rates of autopsy-detected diagnostic errors over time: a systematic review. JAMA. 2003;289:2849–56.
84. Studdert DM, Mello MM, Gawande AA, et al. Claims, errors, and compensation payments in medical malpractice litigation. N Engl J Med. 2006;354:2024–33.
85. Gandhi TK, Kachalia A, Thomas EJ, et al. Missed and delayed diagnoses in the ambulatory setting: a study of closed malpractice claims. Ann Intern Med. 2006;145:488–96.
86. Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. Am J Med. 2008;121:S2–23.
87. Newman-Toker D, Pronovost P. Diagnostic errors—the next frontier for patient safety. JAMA. 2009;301:1062.
88. Zwaan L, De Bruijne MC, Wagner C, et al. A record review on the incidence, consequences and causes of diagnostic adverse events. Arch Intern Med. 2010;170:1015–21.
89. Schiff G, Bates D. Can electronic clinical documentation help prevent diagnostic errors? N Engl J Med. 2010;362:1066–9.
90. Singh H, Giardina T, Meyer A, et al. Types and origins of diagnostic errors in primary care settings. JAMA Int Med. 2013;173:418–25.
91. McDonald K, Matesic B, Contopoulos-Ioannidis D, et al. Patient safety strategies targeted at diagnostic errors- a systematic review. Ann Intern Med. 2013;158:381–90.

92. Society to Improve Diagnosis in Medicine. http://www.improvediagnosis.org. Accessed 26 Jul 2015.
93. National Academies of Sciences, Engineering, and Medicine. Improving diagnosis in health care. Washington, DC: The National Academies Press; 2015.
94. Visualdx. http://www.visualdx.com. Accessed 27 Jul 2015.
95. Papier A. Decision support in dermatology and medicine: history and recent developments. Semin Cutan Med Surg. 2012;31(3):153–9.
96. Goldsmith LA, Papier A. Fighting Babel with precise definitions of knowledge. J Invest Dermatol. 2010;130(11):2527–30.
97. Tleyjeh IM, Nada H, Baddour LM. VisualDx: decision-support software for the diagnosis and management of dermatologic disorders. Clin Infect Dis. 2006;43(9):1177–84.
98. Papier A, Chalmers RJ, Byrnes JA, Goldsmith LA, Dermatology Lexicon Project. Framework for improved communication: the Dermatology Lexicon Project. J Am Acad Dermatol. 2004;50(4):630–4.
99. Papier A, Peres MR, Bobrow M, Bhatia A. The digital imaging system and dermatology. Int J Dermatol. 2000;39(8):561–75.
100. Dxplain. http://www.mghlcs.org/projects/dxplain. Accessed 27 Jul 2015.
101. Greenough A. Help from ISABEL for paediatric diagnoses. Lancet. 2002;360:1259.
102. McKenna C. New online diagnostic tool launched to help doctors. BMJ. 2002;324:1478.
103. Thomas NJ. ISABEL. Crit Care. 2003;7:99–100.
104. Ramnarayan P, Tomlinson A, Rao A, Coren M, Winrow A, Britto J. ISABEL: a web-based differential diagnostic aid for paediatrics: results from an initial performance evaluation. Arch Dis Child. 2003;88:408–13.
105. Ramnarayan P, Kapoor RR, Coren M, et al. Measuring the impact of diagnostic decision support on the quality of decision-making: development of a reliable and valid composite score. J Am Med Inform Assoc. 2003;10:563–72.
106. Berner E. Diagnostic decision support systems: how to determine the gold standard? J Am Med Inform Assoc. 2003;10:608–10.
107. Tsai TL, Fridsma DB, Gatti G. Computer decision support as a source of interpretation error: the case of electrocardiograms. J Am Med Inform Assoc. 2003;10:478–83.
108. Kohn MS, Sun J, Knoop S, Carmeli B, Sow D, Syed-Mahmood T, Rapp W. IBM's health analytics and clinical decision support. IMIA Yearb Med Inform. 2014;9:154–62.
109. Blumenthal D. Stimulating the adoption of health information technology. N Engl J Med. 2009;360(15):1477–149.