Adv in Health Sci Educ (2009) 14:89–106
DOI 10.1007/s10459-009-9186-y

ORIGINAL PAPER

Computer-assisted diagnostic decision support: history, challenges, and possible paths forward

Randolph A. Miller

Received: 14 July 2009 / Accepted: 14 July 2009 / Published online: 12 August 2009

Abstract This paper presents a brief history of computer-assisted diagnosis, including challenges and future directions. Some ideas presented in this article on computer-assisted diagnostic decision support systems (CDDSS) derive from prior work by the author and his colleagues (see list in Acknowledgments) on the INTERNIST-1 and QMR projects. References indicate the original sources of many of these ideas.

Keywords Computers · Decision support systems · Management/trends · Diagnosis · Computer-assisted · Research · User-computer interface

Introduction: definition of diagnosis

Diagnosis is more than the act of associating the name of a disease or syndrome with the findings in a patient case (Miller 1990). Diagnosis is the ongoing process whereby a one elicits from a patient the details (history, symptoms, signs) of how a disease process has unfolded over time, and how that process affects the patient's life situation. Diagnostic evaluation to determine the etiology of a patient's illness often involves sequentially eliciting, over potentially long drawn-out time spans, additional history, symptoms, physical exam signs, laboratory test results, and clinical image interpretations. For some illnesses, diagnosis may entail a "therapeutic trial" to see if the patient responds to a specific intervention in a manner that is characteristic of a specific illness. For example, a physician may administer a 1 week long, low-dose trial of corticosteroids to a patient suspected of having polymyalgia rheumatica, to see if symptoms rapidly resolve. A final aspect of diagnosis involves "diagnosing" the course of an illness over time—e.g., is the disease worsening or improving; is the therapy working as planned; and, has the patient experienced unwanted or intolerable side effects of therapy. Diagnostic evaluation can extend beyond death, through investigative post-mortem autopsy examinations, or delayed

R. A. Miller (✉)
Department of Biomedical Informatics, Eskind Biomedical Library, Vanderbilt University Medical Center, Room B003C, 2209 Garland Avenue, Nashville, TN 37232-8340, USA
e-mail: randolph.a.miller@vanderbilt.edu

genomic investigations as new technologies create new tests. Due to the complex series of activities that must take place over time that comprise "diagnosis", CDDSS can only assist humans with diagnosis tasks undertaken by human clinicians. In the short-term future (i.e., over the next several decades), CDDSS will not replace the role of human clinicians in the diagnostic process (Miller 1990).

A brief history of CDDSS

In human endeavors, insightful pioneers often anticipate future realities. For example, Charles Babbage and Alan Turing foresaw the potential of computing machines and proposed workable mechanisms for building them. Similarly, John von Neumann and Alonzo Church developed computer programming languages before machines existed that could execute such code. The same history holds for CDDSS, where early pioneers envisioned ideas embodied in current systems. Several historical overviews (Collen 1995; Miller 1994) provide a more detailed description of the development of the CDDSS field.

Yearning for diagnostic decision support began at least 2,500 years ago. At that time, an early physician, Hippocrates, wrote in Section I of his Aphorisms, "Life is short, the art long, opportunity fleeting, experience treacherous, judgment difficult." A millennium later, William Shakespeare stated in *Othello* (Act II, Scene 3) that "Men are men; the best sometimes forget" and in *The Merchant of Venice* (Act I, Scene 2), "If to do were as easy as to know what were good to do, chapels had been churches, and poor men's cottages princes' palaces." Shakespeare also acknowledged the potential market for high-quality decision support, writing in 1605, "Good counselors lack no clients" (*Measure for Measure*, Act I, Scene 2). In that same year, Miguel de Cervantes, in his novel *Don Quixote*, commented on the need to evaluate man-made artifacts: "The proof of the pudding is in the eating."

More recently, a series of publications have documented the need for decision support for clinicians. A statement attributed to Sir William Osler circa 1910 was, "half of what we now know about medicine is incorrect, but we unfortunately do not know which half." In 1978, Durack and colleagues physically weighed annual volumes of *Index Medicus* to estimate that published medical knowledge increased exponentially (Durack 1978), and a follow-up assessment by Madlon-Kay a decade later extended this observation (Madlon-Kay 1989). Related to Osler's claim, a 1991 quantitative assessment by Ramsey and colleagues estimated the half-life of knowledge relevant to the practice of Internal Medicine was 5 years (Ramsey et al. 1991). Leigh and colleagues support this observation in 1993 by showing that family practitioners' knowledge declined over time, based on recertification examination performance (Leigh et al. 1993). Numerous studies of clinicians' information needs during inpatient and outpatient care activities indicate that 0.1–5 pertinent but unanswered questions occur per clinician half-day (Osheroff et al. 1991; Gorman and Helfand 1995). A 2009 article by Newman-Toker and Pronovost labeled diagnostic errors as the "next frontier for patient safety" (Newman-Toker and Pronovost 2009).

The modern history of computer-assisted diagnostic decision support systems (Collen 1995; Miller 1994) began in the 1950s, in parallel with the development of digital computing machines. Sentinel publications appeared in *Lancet* by Nash, (Nash 1954) in *JAMA* by Lipkin and Hardy, (Lipkin and Hardy 1958) and in *Science* by Ledley and Lusted (Ledley and Lusted 1959). Over 5,000 CDDSS-related publications have appeared in the peer-reviewed literature (Miller 1994). Those descriptions fall in to three general categories:

(1) Systems for general clinical diagnosis regarding dozens to hundreds of possible disorders. Such systems take as input patient findings and produce a differential diagnosis (and often a suggested workup protocol) as output;
(2) Focused systems for clinical diagnosis that choose among a limited number of possible diagnoses from a clinical subspecialty area. For example, there are systems that interpret the results of pulmonary function tests, arterial blood gas results, or electrocardiogram tracings;
(3) Focused systems that interpret a specific category of images (e.g., pathology slides, cytology smears, digitized xrays, or other imaging modalities) to detect, for example, whether there is evidence of possible malignancy on Pap smear slides (Miller 1994).

This article concerns only the first of those three categories—general clinical diagnostic systems.

In their 1959 *Science* article (Ledley and Lusted 1959). Ledley and Lusted noted that physicians have imperfect self-knowledge of their own diagnostic problem solving methods. Newell and Simon's classic 1972 book, *Human Problem Solving*, experimentally verified a more general version of that observation—that humans gain knowledge and expertise by "compiling" what they read and experience into "chunks" that they can apply as a form of rapid pattern recognition (Newell and Simon 1972). Ledley and Lusted suggested that protocol analyses could contribute substantially to the understanding of diagnostic reasoning. They stated that computers could assist with two important aspects of clinical diagnosis—deductive logic (as embodied in set theory and Boolean algebra in computer systems) and probabilistic reasoning (as embodied in Bayes' rule on computers). They predicted that, based on application of Bayes' rule, CDDSS using decision-analytic approaches would be possible.

In 1960–1961, Warner and colleagues at LDS Hospital in Salt Lake City developed the first operational Bayesian CDDSS, for the diagnosis of congenital heart diseases based on history, physical exam, and cardiac catheterization findings (Warner et al. 1961). They noted the importance of assuming that diseases and findings occurred independently of one another to simplify Bayesian calculations. They demonstrated how case findings correspondingly could be modified to eliminate redundant, non-independent findings before analyses. They derived the probabilities required for their system's Bayesian calculations from both literature review and their own series of 1,000 cardiac catheterizations. They observed that their system was very sensitive to both false positive findings in patient cases, and to errors in the system's database. Finally, they pointed out the need for an independent "gold standard" to judge the diagnostic performance of their CDDSS.

The systems based on Bayes' rule have evolved from an early model, whereby the user entered all known findings initially, in one session, and received a differential diagnosis—as pioneered by Warner et al.—to a later model developed by Gorry and Barnett (1968) that included a Bayesian sequential questioning strategy to serially elicit findings and refine the differential diagnosis. From the late 1960s through the 1990s, de Dombal and colleagues at Leeds University in England organized widespread UK and European clinical trials of a Bayesian system for the diagnosis of acute abdominal pain (de Dombal et al. 1971; Horrocks et al. 1972; Adams et al. 1986). A number of groups further explored Bayesian diagnostic approaches by relaxing the requirements for the independence assumption (Fryback 1978), exploring "mixed" Bayesian and non-Bayesian models (Warner's Iliad CDDSS in the mid-1980s; Warner et al. 1987), and development of a more general approach using Bayesian Belief Networks to model all conditional dependencies in a diagnostic system (pioneered by Pearl (1987) and Cooper (1986), among many others).

A number of other approaches to computer-assisted diagnosis include statistical clustering models (e.g., Nordyke et al. 1971), which have evolved into present-day CDDSS approaches such as support vector machine systems (e.g., Statnikov et al. 2005); branching logic ("20 questions") CDDSS of the sort pioneered by Bleich and colleagues for diagnosis of acid-base disorders in the 1960s (Bleich 1969); semiquantitative or quantitative physiological models, of the sort initially pioneered by Guyton and colleagues in Mississippi in the 1960s, and later expanded to include multi-tiered CDDSS explanatory models of the sort developed by Patil and Szolovits for acid-base disorders in the early 1980s (Szolovits et al. 1988); heuristic CDDSS based on criteria tables, with major and minor features designated to establish a diagnosis when present in certain combinations (Lipkin and Hardy 1958; Lindbeg et al. 1968; Blois et al. 1981; Kingsland et al. 1982, 1983; Porter et al. 1988)—akin to the Jones Criteria used for the diagnosis of acute rheumatic fever (Shiffman 1995); and, rule-based diagnostic systems, such as Shortliffe's pioneering MYCIN project developed from 1974 to 1976 (Shortliffe 1976).

A final category of CDDSS merits special attention: heuristic (empirical, "rule of thumb") approaches using symbolic reasoning. The pioneer G. Anthony Gorry, in a 1968 article, laid the foundation for the next half-century of work in this area (Gorry 1968). Gorry sketched general principles required for heuristic diagnostic expert systems. His schemata included the need to begin with a formal definition of the diagnostic problem; that essential heuristic CDDSS components included a general-purpose inference function that could generate diagnoses from observed findings, a generic test-selection function that could dynamically identify the best test to order at any point in a diagnostic evaluation; and, a generic pattern-sorting function to determine which diagnoses are competitors, where only one of the set is likely to be present (i.e., fall within the same "problem area"—such as potential heart diseases that a patient might have), and which are non-competitors, and potentially all present (i.e., from different problem areas, such as a heart disease, a liver disease, and a kidney disease). Gorry described important differences among the information-entropy value of a test, the economic cost of the test, and the morbidity/mortality risk of performing a test as factors that determined which test would be optimal to select next. He further noted that CDDSS should consider the costs of misdiagnosis of life-threatening or disabling disorders, re-emphasized the early observation of Warner and colleagues regarding the potential negative influence of "red-herring" findings in case descriptions, and described the complexities introduced by patients with multiple concurrent diagnoses, where seeking the single best diagnosis compatible with all of the patients' findings could at times be counterproductive. Finally, Gorry suggested that the knowledge bases underlying CDDSS might be used to generate educational patient simulations (Gorry 1968).

Conceptual descendants of Gorry's schemata-based heuristic CDDSS include: the present illness program (PIP), developed by Pauker et al. (1976) for taking a history of the present illness in patients with renal disease; INTERNIST-1 (developed by Myers, Pople, and Miller) and Quick Medical Reference (QMR), originally developed by Miller, Masarie, and Myers for diagnosis in general internal medicine (as described in detail below; Miller 1984, 1990, 1997; Pople et al. 1975; Miller et al. 1982, 1986a, b; Pople 1982; Giuse et al. 1989a, b, 1990, 1993a, b, c, 1995; Masarie et al. 1985; Bankowitz et al. 1989a, b; Miller and Masarie 1989, 1990, 1992; Giuse and Bankowitz 1993; Aliferis et al. 1996; Miller and Schaffner 1982; Parker and Miller 1989); DXplain, developed by Barnett and colleagues at Massachusetts General Hospital for diagnosis in general medicine (Hupp et al. 1986; Barnett et al. 1987; Feldman and Bartlett 1991); Iliad, which combined Bayesian and symbolic reasoning in a system developed by Homer Warner and colleagues at LDS Hospital in Utah (Warner et al. 1987; Warner 1989; Lau and Warner 1992); and, a number of more recent systems, including ISABEL (Ramnarayan et al. 2004, 2003, 2006), a commercially available system.

The INTERNIST-1/Quick Medical Reference project as an example of CDDSS; INTERNIST-1 Project, 1972–1983

From 1955 to 1970, Jack D. Myers, MD, served as Chairman of the Department of Medicine at the University of Pittsburgh. During his career, he established a reputation as one of the most discerning diagnosticians in American internal medicine. When he voluntarily stepped down after serving 15 years as Chairman of Medicine, he sought out a collaborator to work on computer-assisted diagnosis. Myers was able to engage Harry E. Pople, Jr., who had worked on computer programming as an undergraduate at MIT, and then in 1968 obtained his PhD in Industrial Engineering at Carnegie Mellon University (CMU) before joining the faculty of the School of Business at the University of Pittsburgh. In March 1973, The author (Randolph Miller) joined the project as a second-year medical student, in the role of volunteer programmer. The INTERNIST-1 program that they developed followed an empirical approach to diagnosis.

The goals of the INTERNIST-1 project were to develop computer algorithms and a clinical diagnostic knowledge base (KB) that could support expert consultations for diagnosis in general internal medicine; to create a program which accepted as input the patient's history, physical exam, and laboratory data; to have the program produce output consisting of either concluded diagnoses or a "working" differential diagnosis; to have the program able to lead physicians through cost-effective patient "work-ups"; to develop a KB for clinical diagnosis that comprised an academic, evidence-based repository of what was known through the peer-reviewed literature about the findings of each given disease; and, to maintain the KB through testing and feedback using actual patient cases with carefully established diagnoses.

Pople used the protocol analysis techniques developed by Simon at Carnegie Mellon University to analyze Myer's logic as he "thought aloud" while solving challenging diagnostic cases. In parallel with diagnostic algorithms that Pople developed in the 1973–1974 time frame (Pople et al. 1975; Miller et al. 1982; Pople 1982), Myers developed the INTERINST-1 KB in conjunction with medical students who conducted extensive literature reviews on diseases of interest to them, and presented their results to Myers. Miller wrote the INTERNIST-1 KB Editor program in a combination of LISP and Assembly Language. The latter was required to overcome computer memory limitations in the early LISP environment. Myers' decision to build the knowledge base using the biomedical literature as the "gold standard" was critical; the common model for building expert systems at the time was to have a knowledge engineer (computer scientist) pair with a domain expert (in this case, a highly regarded physician), and to "debrief" the domain expert to directly record his or her knowledge as the basis for the expert system. Unfortunately, such an approach is not scientifically reproducible, as use of different domain experts with that model would produce different knowledge bases. A later study by Nunzia Giuse and colleagues established that the approach preferred by Myers was indeed scientifically reproducible—i.e., different experts in different settings, given proper background and tools, could independently create essentially the same disease profiles, given the peer-reviewed literature as a source (Giuse et al. 1993c).

The INTERNIST-1 diagnostic algorithms followed the general form for expert system schemata previously outlined by Gorry, with a number of important innovations. First, the knowledge base was literature-derived, and addressed the question, "what has been reliably and verifiably reported to occur in patients with this disease?" (Miller et al. 1982) Second, a separate compartment of the knowledge base kept track of known dependency relationships among findings and diseases ("properties"; Masarie et al. 1985). As Myers liked to say, the function of properties was "to keep the egg off the face of the computer". Properties encoded that only females could become pregnant, that persons with a history of cholecystectomy could not later develop acute cholecystitis, that persons with atrial fibrillation should not be examined in an attempt to find an S4 atrial gallop, and that if one knows that a patient is older than 55 years of age, then one can conclude that the patient is not younger than 26. Properties prevented the system from eliciting multiple redundant findings during case workup mode by suppressing question asking for such findings. Third, in the KB, associated with each finding in each disease profile were two numbers—an evoking strength (positive predictive value) on a 0–5 integer scale, and a frequency (sensitivity) on a 1–5 integer scale. Figure 1 provides an excerpt of the INTERNIST-1/QMR KB profile for "Aortic Valvular Stenosis".

The legend of Fig. 1 recaps the meaning of evoking strength and frequency numbers (Miller et al. 1982). From 1973 to 1999, an estimated 40 person-years of effort went into KB development for approximately 650 disease profiles of disorders seen by general internists. The KB included a number of surgical and gynaecological disorders that internists frequently encounter.

To obtain an INTERNIST-1 diagnostic consultation required 45–90 min of time for users to abstract the case, enter all pertinent initial positive and negative findings, and for the system to analyze the case. Figure 2 illustrates the process of INTERNIST-1 case analysis (see Miller et al. 1982 for details).

The system generated a ranked, scored differential diagnosis list, sorted it into competitors and non-competitors for the topmost diagnosis, temporarily discarded the non-competitors, and examined the remaining list to determine: (a) whether the topmost diagnosis was sufficiently higher in score than its competitors to be able to "conclude" it, or (b) which questioning mode to employ—PURSUING if the topmost diagnosis was near the threshold for conclusion, DISCRIMINATE if there were a small number of competitors at the top of the list, and RULE OUT if there were a large number of competitors close in score at the top of the list The INTERNIST-1 question-generating algorithm would look for inexpensive discriminating questions (e.g., history and physical exam findings) before escalating to more expensive and invasive forms of test-seeking. For example, at the bottom of Fig. 2, a history of diabetes mellitus would favor myocardial infarction over aortic dissection, whereas a family history of Marfan's syndrome would favor the latter diagnosis. After asking a limited number of questions, the system would re-analyze the case "from scratch", using the additional information gained from questioning to recalculate scores and competitor/non-competitor partitions. The latter strategy resulted in seemingly intelligent behavior by the system—when the questions related to a set of cardiovascular "competitor" diagnoses were negative (findings absent in case), the scores of those diseases would fall, and a new topmost diagnosis, e.g., a liver disease, might emerge. Its competitors would be other liver diseases, so it appeared to a casual observer that the system had intelligently shifted its focus to pursue another line of reasoning. In this somewhat random manner, INTERNIST-1 would explore sets of competitor diagnoses until it found an area in which it could make a diagnosis. This often led to rapid resolution of other diagnoses, as described in (Miller et al. 1982), since findings "explained" by the concluded diagnosis were removed from the case description, and the system then recycled using any remaining unexplained findings, until all significant case findings were

**** ABRIDGED from ORIGINAL LIST with 92 entries ****

Past Medical History and Demographics (abridged) ...
- 0 1 Age 16 to 25
- 0 2 Age 26 to 55
- 0 3 Age Gtr Than 55
- 2 2 Angina Pectoris Hx
- 1 2 Heart Disease Family Hx
- 2 2 Rheumatic Fever Hx

Symptoms of Current Illness (abridged) ...
- 2 3 Dyspnea Exertional
- 1 2 Chest Pain Substernal At Rest
- 2 2 Syncope or Syncope Recent Hx

Findings on Physical Examination (abridged) ...
- 2 4 Heart Impulse Apical Forceful Localized
- 3 4 Heart Murmur Systolic Ejection Second Right Interspace
- 3 3 Heart Murmur Systolic Ejection Second Right Interspace Transmitted To Neck
- 2 3 Heart Percussion Left Border Lateral Displacement
- 3 3 Heart Sound(s) A2 Decreasedng Valsalva
- 2 2 Heart Murmur Diastolic Apical Immediately After S2
- 1 2 Heart Murmur Holosystolic Apical
- 2 2 Heart Murmur Systolic Apical
- 2 2 Heart Murmur Systolic Ejection Left Sternal Border
- 2 2 Heart Murmur Systolic Ejection Second Right Interspace Transmitted To Apex
- 4 2 Heart Thrill Systolic Second Right Interspace
- 2 2 Heart Thrill Systolic Suprasternal Notch
- 4 2 Pulse Arterial Anacrotic
- 3 3 Pulse Arterial Plateau
- 2 2 Pulse Pressure Narrow

Simple Inexpensive Laboratory Findings (abridged) ...
- 1 2 EKG Atrial Fibrillation
- 2 4 EKG Left Axis Deviation
- 2 4 EKG Left Ventricular Hypertrophy
- 4 2 Heart Xray Aortic Valve Calcified

Intermediate Mildly Expensive or Invasive Findings (abridged) ...
- 2 2 Aorta Ultrasonography Aortic Root Diameter Increased
- 2 3 Heart Echocardiography Aortic Valve Bicuspid
- 3 4 Heart Echocardiography Aortic Valve Leaflet <s> Diffuse Thickening

Moderate to Very Expensive or Invasive Findings (abridged) ...
- 4 4 Heart Angiocardiography Aortic Stenosis Valvular
- 3 4 Heart Catheterization Left Ventricle Aorta Pressure Gradient Increased

Linked (Associated) Disorders (abridged) ...
- 2 2 Causes Cardiac Failure Left Chronic Congestive
- 2 2 Causes Left Ventricular Failure Acute
- 2 2 Predisposes to Endocarditis Subacute Infective Left Heart
- 2 1 Predisposes to Endocarditis Acute Infective Left Heart
- 2 2 Coincides with Aortic Regurgitation Chronic
- 2 2 Preceded by Rheumatic Carditis Acute

N.B. The two numbers on the left margin beside each finding are an evoking strength (PPV on 0-5 scale with 0 nonspecific and 5 pathognomonic) and on the right, a frequency (sensitivity on 1-5 scale, where 1=rare, 3=half, 5=finding always occurs in patients with disease)

Also note: Findings of associated disorders, such as Acute Rheumatic Fever and Chronic Congestive Left Heart Failure, are profiled separately and not listed here, as designated by "Links" below – See reference [5]

Fig. 1 (ABRIDGED) *INTERNIST-1/QMR disease profile: Aortic valvular stenosis*

explained, or the system ran out of viable questions with a remaining, unresolved differential diagnosis.

Important lessons learned during the INTERNIST-1 project included: (1) develop the knowledge base for the system independently from the diagnostic algorithms, using an external gold standard, such as the peer-reviewed literature; (2) use feedback from actual patient case analyses to determine if flaws in the knowledge base or in the diagnostic algorithms led to any failures; (3) only change the KB by examining the relevant literature to see if the initial literature review had been incomplete or inaccurate, and avoid changing the KB arbitrarily just to improve performance in an individual case. It is important to maintain a battery of "standard test reagents"—a set of cases with known diagnoses that are re-analyzed by the system periodically to make certain that the performance of the system does not degrade or drift over time. (4) Clinicians by nature are often inexact in

```
+SEX Male
+AGE Gtr Than 55
+ABDOMEN Pain Epigastrium
+ABDOMEN Pain Severe
+UNCONSCIOUSNESS Recent Hx
+HYPERTENSION Hx
+MYOCARDIAL Infarction Hx
+ANGINA Pectoris Hx
+HEART Catheterization Recent Hx
+CORONARY Arteriography Fixed Luminal Narrowing 70 Percent Or Gtr
+HEART Angiocardiography Left Ventricle Adynamic Area <S>
+HEART Surgery Recent Hx
+PRESSURE Arterial Diastolic Gtr Than 125
+DYSPNEA At Rest
+BOWEL Sound <S> Decreased          USER ENTRY OF CASE FINDINGS

+GO                    "GO" INDICATES FINISHED ENTERING FINDINGS

CONSIDERING: SEX Male, AGE Gtr Than 55, ABDOMEN Pain
Epigastrium, ABDOMEN Pain Severe, UNCONSCIOUSNESS
Recent Hx, HYPERTENSION Hx, MYOCARDIAL Infarction Hx,
ANGINA Pectoris Hx, HEART Catheterization Recent Hx,
HEART Surgery Recent Hx, PRESSURE Arterial Diastolic
Gtr Than 125, DYSPNEA At Rest

                    INTERNIST-1 SYSTEM INDICATES WHICH FINDINGS
                     ARE CONSISTENT WITH ITS TOPMOST DIAGNOSIS

DISCRIMINATE: AORTIC DISSECTION, MYOCARDIAL INFARCTION ACUTE

           INTERNIST-1 QUESTION-GENERATING MODE (DISCRIMINATE)
                       AND CURRENT DIFFERENTIAL DIAGNOSIS LIST

DIABETES MELLITUS HX?

MARFANS SYNDROME FAMILY HX?

MYOCARDIAL INFARCTION FAMILY HX?

      INTERNIST-1 QUESTIONS FOR USER TO ANSWER TO WORK UP CASE
```

Fig. 2 (ABRIDGED) *INTERNIST-1 case analysis for N Engl Med V324 P527 1991*

their thinking. By forcing them to use the carefully crafted set of patient descriptors created by Myers, who was an expert diagnostician, the system already gained an advantage by forcing the clinician-user to be more precise in characterizing the patient's illness, thus leading to improved diagnostic accuracy; (5) It is important to have a senior, well-respected, truly expert clinician associated with any clinical informatics project—for improved quality of knowledge base and algorithms during development, for credibility once the system is demonstrated to the public, and for perseverance and insight during evolution and maintenance of the project. As an extra benefit, such individuals are typically senior enough to not be caught up in earning a reputation (they have already done so), or in achieving tenure, and can therefore focus more clearly on the project. (6) It is critical to have a scientifically reproducible method of knowledge base construction and maintenance, based on a commonly accessible, high-quality, "gold standard" source, such as the peer-reviewed literature. (7) It is possible to construct diagnostic algorithms and associated knowledge bases that can rival capable human diagnosticians in their ability to advise clinicians facing diagnostic challenges (Miller et al. 1982; Bankowitz et al. 1989a).

Quick Medical Reference (QMR) project, 1984–2002

The INTERNIST-1 approach assumed that the clinician had a diagnostic problem, and should at that point relinquish all control of further diagnostic evaluation to the consultant "Greek Oracle" system, which would return the ultimate true diagnoses, based on asking the clinician-user to work up the patient in a specified manner. There were several flaws in this assumption (Miller and Masarie 1990). First, when clinicians encounter diagnostic problems, it is not because they have forgotten everything they have learned about diagnosis, or lost the ability to conduct a diagnostic evaluation. It is more likely that a specific step of a multi-step diagnostic process has become puzzling to the clinician in a specific case. Typically, the clinician will seek advice from a respected colleague, do a MEDLINE search, cast a wide net of diagnostic tests to see what turns up, or assume a posture of "watchful waiting" until more obvious signs of the illness develop. The rate-limiting problem may be simple—e.g., the clinician does not know or remember the differential diagnosis of a rare and unusual finding in the case, or the clinician is not familiar with a very rare disorder that is a candidate diagnosis for the patient, or there may be a small number of vague findings troubling the patient that do not point to any specific diagnosis, or there may be an overwhelming number of positive findings that the clinician has difficulty sorting out. Myers and Fred E. (Chip) Masarie, Jr., MD worked with Miller to create QMR and its associated knowledge base as a diagnostic decision support system that allowed the clinician user more flexibility than the "Greek Oracle" model of INTERNIST-1.

The goals in developing QMR (Miller et al. 1986a, b; Bankowitz et al. 1989a, b; Giuse et al. 1989a, b, 1990, 1993a, b, c, 1995; Miller and Masarie 1990, 1992; Giuse and Bankowitz 1993; Aliferis et al. 1996; Miller 1997) were to: (1) make maximal use of the existing INTERNIST-1 KB in order to create a "diagnostic tool kit" that a clinician might consult for a few seconds or a few minutes, to overcome the step in diagnosis that caused the problem; (2) at all times recognize the expertise of the clinician-user, in the role of the system's "pilot" who determined where to go and how to get there; (3) emphasize support of real-world diagnostic decision-making by physicians, rather than developing a tour-de-force expert consultant with maximum utility for rare, complex challenging cases; and, (4) build a system that could augment the diagnostic performance of the unaided clinician-user.

The work of Drs. Nunzia and Dario Giuse led to their development of a new QMR knowledge acquisition tool (QMR-KAT), which enforced principles of rigor and consistency while assisting clinicians engaged in knowledge base construction (Giuse et al. 1989a, b, 1990, 1993a, b, c, 1995; Giuse and Bankowitz 1993). The QMR-KAT enabled KB builders to enter audit trails for multiple references supporting how it is known that a given finding happens in a given illness. QMR-KAT further allowed KB builders to provide several forms of frequency (sensitivity) estimates of the finding for the disease on a reference-by-reference basis, and to provide an overall figure of merit for how pertinent and reliable a given article was in documenting the occurrence of the finding in the disease.

A sample, abbreviated QMR-KAT disease profile for Perinephric Abscess appears in Fig. 3.

The QMR program operated at three levels of diagnostic support. First, QMR served as a simple electronic textbook that could display disease profiles or findings' differential diagnoses sorted in various ways (by sensitivity, predictive value, clinical importance of

**** ABRIDGED from ORIGINAL LIST with over 150 entries ****

- 0 3 AGE 26 TO 55
[1]3 18 pts, range 35-104, median 63; p. 72
[5]3 67% of 103 pts were 26-55 yrs old, 1925-1940
[13]3 44/74 cases at Bellvue, 1920-1930
[70]3 15/20 in 1929 author's series
[82]3 Majority of 55 pts "between 30 and 40" (1931)
[91]4 Average age 54 in 1979-83 series of 15 pts

- 1 2 DIABETES MELLITUS HX
[1]3 4/18 cases had hx of diabetes mellitus
[13]2 3/83 cases, 1920-1930
[88]4 10/71 cases had diabetes, 1953-1965
[91]2 3/15 cases had diabetes as predisposing factor in 1979-83 series
[95]2 Case report

- 1 2 URINARY CALCULUS HX
[12]2 Case report
[40]2 Case report in author's series
[70]2 Case report in author's series
[58]2 Case 3 in author's series

- 2 4 BACK PAIN COSTOVERTEBRAL ANGLE <S>
[1]2 12/18 had CVA tenderness, p.73
[5]3 96 of 117 cases, 1925-1940
[13]2 64/83 had "loin pain"
[103]4 25/26 patients had flank pain
[101]4 38/52 pts had unilateral flank pain (73%)
[70]3 20/21 of author's cases, 1929

- 2 2 BACK INSPECTION UNILATERAL INFRACOSTAL FULLNESS AND/OR OBLITERATION OF ILIOCOSTAL CURVE
[30]2 Documented photographically, several cases
[70]2 Tumefaction mentioned in several cases of author's series (1929)
[89]3 "Waistline is obliterated" base on series of 103 cases

- 2 4 BACK TENDERNESS COSTOVERTEBRAL ANGLE <S>
[1]3 12/18 had CVA tenderness, p.73
[5]2 Tenderness in loin, 54/117 cases, 1925-1940
[101]4 73% current series; two-thirds had either flank mass or unilateral tenderness on literature review
[25]3 38/49 had CVA or loin tenderness

- 2 4 ABDOMEN COMPUTERIZED TOMOGRAPHY GEROTAS FASCIA THICKENED
[11]4 10/11 patients in author's series
[27]2 focal ipsilateral thickening of Gerota's fascia

- 3 5 ABDOMEN COMPUTERIZED TOMOGRAPHY PERINEPHRIC FLUID COLLECTION
[11]4 11/11 cases in author's series
[110]2 Trauma main other cause; might make evoking strength 4

^ 1 1 RENAL TUBERCULOSIS (pdis) PERINEPHRIC ABSCESS
[41]3 10/106 Mayo cases due to TB, 1914-1924
[91]2 1/15 cases due to AFB in 1979-83 series

Fig. 3 (ABRIDGED) *QMR-KAT disease profile: perinephric abscess* findings, etc.). Second, QMR functioned as an intermediate combinatorial tool that could quickly provide, in a few seconds, the differential diagnosis of a small number of "key findings" from a case (Miller and Masarie 1992), as illustrated in Fig. 4; generate cost-effective "work-up" suggestions for a differential diagnosis entered by the user (as opposed to generated by QMR); and, show closely related diseases to a given disease (based on degree of overlap of important findings in both disease profiles.

At the third level, QMR provided the same sort of diagnostic consultation capabilities as INTERNIST-1, with the exception that the steps in the process were under greater control of the user (e.g., the user could decide when to generate questions, or designate that a certain diagnosis be treated as if it were the topmost diagnosis for generating competitors). A prospective evaluation of QMR during actual clinical practice demonstrated that clinicians using QMR on inpatients with challenging, unknown diagnoses were better able than the academic ward team caring for the patient to identify the correct diagnosis, as determined by diagnostic evaluation over the ensuing 6 months after consultation (Bankowitz et al. 1989a, b).

Lessons learned from INTERNIST-1 and QMR projects about CDDSS evaluation

In the course of evaluating INTERNIST-1 and QMR, project members have developed the following perspectives, which supplement the findings of formal external evaluations (Berner et al. 1994; Friedman and Wyatt 1997; Friedman et al. 1999; Berner et al. 1999). A "first principle" for CDDSS evaluation (Miller 1996) is that clinicians, not systems or evaluators, discover, characterize, and attempt to solve clinical diagnostic problems. As a result, CDDSS provide benefit only when they assist clinician-users to solve problems that they cannot solve on their own. It follows that in vivo evaluations of CDDSS, on actual patients at a time when no one knows the correct diagnosis, is the most definitive way to determine CDDSS benefit. The ability of a system to solve "artificial" cases presented to it

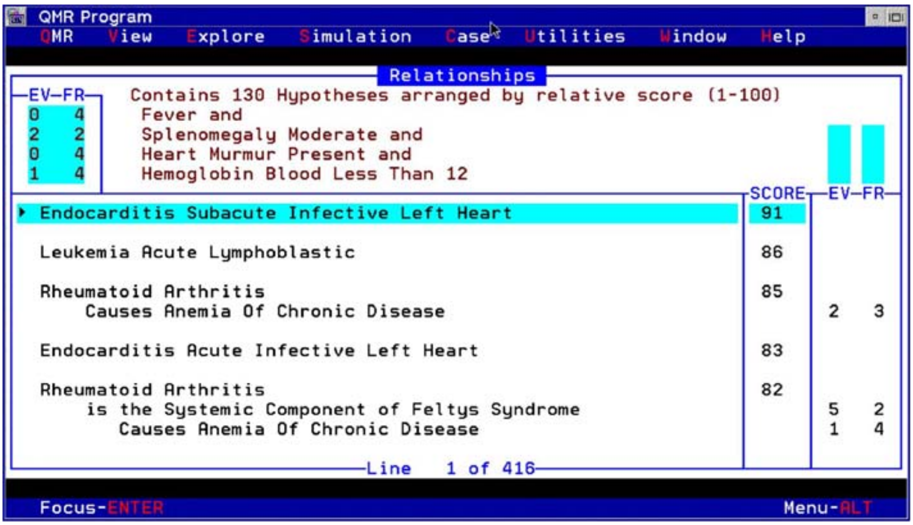

Fig. 4 (ABRIDGED) *QMR Relationships function for: fever, heart, murmur, splenomegaly, and anemia*

in a manner "better" than clinicians who attempt to solve the same "paper-based scenario" case description in a classroom setting ("in vitro" evaluation) is a useful waypoint along the CDDSS development and evaluation spectrum, but that approach is far less definitive and insightful than is in vivo evaluation. Thus, systems should not undergo definitive evaluations as freestanding "Greek oracles" based on their ability to solve clinical problems autonomously. The CDDSS should be measured on their ability to help clinicians solve the clinicians' own real-world problems during patient care. The ultimate unit of evaluation should be whether the user plus the system performs better than the unaided user (Miller 1996).

It is important to determine, before undertaking an evaluation, what the criteria for efficacy of the CDDSS should be (Miller 1996). For example, are the criteria for "successful" system performance similar to what clinical practitioners would use or require during actual patient care delivery? Determining the meaning of "diagnostic benefit" and how to measure it can be non-trivial. Randomized controlled trials typically include careful entry criteria and exclusion criteria to determine patient eligibility. Similarly, for CDDSS evaluation, there must be objective criteria for how to establish a "gold standard" diagnosis in a manner independent from the input data stream used by the CDDSS (i.e., there must be at least one "definitive test" or set of criteria that are established outside of the context of the data input into the system, though data up to that definitive test can be used in CDDSS analysis). For retrospective studies, one might develop criteria for what sorts of "clinching" diagnostic case data to withhold from CDDSS case analysis. However, for prospective studies, there must be a plan for how patients will be followed to establish a diagnosis, and what the inclusion and exclusion criteria for concluding a diagnosis are. Simply using ICD-9 codes for diagnoses that were assigned during routine care, in either the inpatient or outpatient environments, is not adequate, since those codes are used for billing purposes and are prone to biases and inaccuracies, and do not well capture the level of disease activity during the admission.

One potential measure of both human and CDDSS diagnostic performance might be therapeutic benefit—advancing knowledge of the patient case sufficiently to establish a correct therapy, for example. Under such circumstances, if a CDDSS mistakenly proposes a diagnosis of pulmonary histoplasmosis, and the actual diagnosis is pulmonary cryptococcosis, the diagnostic error is less significant, since the same anti-fungal agents are effective for therapy in both illnesses. A misdiagnosis of adenocarcinoma of the lung under such circumstances would be much more serious, because the treatment would be radically different than for a fungal infection.

Another lesson learned during the INTERNIST-1 and QMR projects was that using a CDDSS KB to generate "simulated" patients for diagnostic practice by trainees and practitioners (Parker and Miller 1989)—as was originally suggested by Gorry (1968)—can provide KB developers with another valuable form of feedback. If simulated cases developed using a CDDSS KB do not seem to be like "real" patients in ways that clinicians can articulate, then underlying defects in KB representations may be exposed, such as inability to represent severity of findings or severity of illness, or time course of illness, in a substantive manner (Miller et al. 1982). In addition to simulation, there are a number of other useful educational aspects of CDDSS KBs for biomedical and informatics trainees (Miller and Schaffner 1982; Miller and Masarie 1989; Giuse et al. 1989a, b; Parker and Miller 1989).

Another important lesson in evaluating CDDSS is that one must carefully study the intended purpose of CDDSS developers in creating a specific system, and examine the coverage of its KB (Miller 1996). Clinical end-users are likely to be insensitive to whether the possibly unknown diagnosis of their patients are included in the system or part of its intended use. Thus, system evaluation should include determination of how the system performs on cases outside the scope of the systems' intended use (such studies should be carried out in a manner that protects patients). It is far worse if a CDDSS always offers an incorrect diagnosis under such circumstances, than if it always "fails" to conclude any diagnosis (given that the correct diagnosis is outside the scope of the system). The consequences of misuse should be evaluated. For example, de Dombal's system for diagnosis of acute abdominal pain was designed for use in hospital emergency departments, (de Dombal et al. 1971; Horrocks et al. 1972; Adams et al. 1986) to help screen for patients who required surgical interventions (such as appendectomy or cholecystectomy). One diagnostic category employed by the system was "non-specific abdominal pain", which was intended for labeling non-surgical causes of pain. A patient presenting to the emergency department with colicky abdominal pain and a dark line along the gums in the mouth might be correctly labeled as having "non-specific abdominal pain", but the lead poisoning that caused the findings is a serious and treatable illness. The purpose of the system was not to detect lead poisoning, it was to determine if the patient required surgery. System users and evaluators must be cognizant of such limitations.

Finally, there are ethical and legal issues related to the use of computer programs in clinical practice (Miller et al. 1985). One should not allow use of such systems without proper training. There exists a cognitive dissonance between what a clinician expects to be able to obtain from an unknown CDDSS, and what the CDDSS can actually provide as assistance. With proper training, clinicians temper unrealistic expectations, and are often surprised at some of the capabilities the systems can provide. As noted above, one must make sure that a CDDSS is used for the purposes for which it was intended, and not for out-of-scope applications. Persons with limited ability to override mistaken advice from a CDDSS should not be permitted to use the CDDSS for serious purposes (such as for live patient care or for denying insurance claims). Users must ascertain the quality and currency of any CDDSS KB before accepting advice from the system. A system "validated" for clinical use in 1 year is unlikely to be of value 20 years later without updates to the KB, since clinical practice changes rapidly (Miller et al. 1985). System assessment and validation should be an ongoing process, as described above.

Problems and challenges facing CDDSS

There are a number of problems facing implementers of CDDSS that have been understood for decades (see e.g., Miller 1984). Despite many years of study, as the other articles in this supplement illustrate, what is known about human diagnostic processes remains limited and controversial. As a result, there is no well-understood "roadmap" of how humans go about diagnosis that could easily be converted into a widely accepted and applied CDDSS application.

Some forms of CDDSS implementation impediments and resistance are generic to all forms of clinical decision support (CDS). A key source of CDS resistance follows from clinical care delivery occurring in chaotic environments, with unexpected interruptions. Clinical dilemmas are often encountered during times of incomplete information. David Bates concisely outlined many CDS implementation obstacles in his classic 2003 article on the ten commandments for effective clinical decision support (Bates et al. 2003): speed of response is critical for busy clinicians; anticipate CDS needs and deliver answers in real time; adapt CDS to fit into the user's workflows; usability matters a lot; clinicians strongly resist suggestions not to carry out an action, unless an easy way to carry out an alternative is offered; simple interventions work best; monitor impact, get feedback, and respond; and, manage and maintain knowledge-based systems (Bates et al. 2003).

Another CDDSS implementation problem is that the advice that a CDDSS offers to highly trained, knowledgeable clinicians must be of sufficient quality to merit the clinicians' trust and respect. To paraphrase Yogi Berra, "To deliver expert clinical diagnostic advice, you have to deliver expert clinical diagnostic advice." Despite widespread promulgation and general acceptance of "evidence-based clinical practice" as being desirable, there currently exists no centralized high-quality repository of ready-to-use clinical diagnostic knowledge content—either for human or CDDSS consumption. The biomedical literature is such a resource, but it has been likened to trying to drink from a fire hose. As a result, CDDSS developers often recapitulate previous content development work, with unnecessary variation in both quality and format of the KBs that underlie each CDDSS.

The future of CDDSS

For future CDDSS to gain widespread adoption and support, they must comply with Bates' Ten Commandments. Future CDDSS must become better integrated into clinicians' workflows. Consultations involving CDDSS should take seconds to at most a few minutes, not hours, to complete. Ideally, the advice from CDDSS should be given in actionable form. For example, ordering CDDSS-recommended tests should only require a few additional clicks via an integrated care provider order entry (CPOE) system. Similarly, CDSSS should obtain their inputs more directly, either from fully automated natural language processing abstraction of a patient's electronic medical record (EMR), or more realistically, from clinicians highlighting, or clicking on, those findings in the EMR interface that they would like to transfer to the CDDSS for consideration, assuming that such a process could be accomplished objectively.

In the current age of the Internet, MEDLINE®, the Cochrane Database of Clinical Trials®, and Google®, it is unfortunate that there is no national repository of authoritative clinical diagnostic knowledge. Ideally, governmental agencies, private foundations, clinical specialty organizations, and academic medical centers could collaborate to build and maintain an unbiased, respected repository of diagnostic knowledge. The cost of doing so would probably be far less than the savings in healthcare expenditures that reliable diagnostic KB availability would generate when used during care. Such a CDDSS knowledge repository could support clinical education programs for students in medicine, nursing, pharmacy, and other health professions; support a wide variety of non-commercial and commercial CDDSS applications; and stimulate biomedical research by highlighting areas where inadequate knowledge exists about various diseases.

Finally, the healthcare informatics and clinical specialty communities should establish standards for how to evaluate CDDSS performance. In parallel, methods to determine and monitor the quality of their underlying KBs should be agreed upon and applied. In the first issue of the Journal of the American Informatics Association, Dr. William Stead and the then-current members of the National Library of Medicine BLRC Study Section (the study section which reviewed and funded many of the CDDSS development and evaluation projects) issued a paper that illustrated how CDS systems should be evaluated based on the stage of maturity of their development (Stead et al. 1994). Publications by highly respected evaluators in our field, as noted above, have added to our knowledge in this area. Yet, there are still CDDSS developers who do inadequate, informal, in vitro type evaluations to justify promotion of their systems for use in clinical practice. It is problematic to try to develop "certification" exams for such systems, or close regulation by agencies such as the US FDA (Berner et al. 1994), since even minor day-to-day changes in KB or algorithms could dramatically alter system performance. Nevertheless, there should be widely agreed upon standards for the types of in vitro testing (Miller and Gardner 1997) and ultimately in vivo testing (e.g., randomized controlled trials) that should be conducted prior to any widespread CDDSS dissemination (Friedman and Wyatt 1997).

As clinicians more frequent use electronic medical records and computerized physician order entry systems in practice, the ability to integrate CDDSS systems into clinical workflows will increase. Given definitive, unbiased national sources of diagnostic knowledge, and agreed upon procedures for clinically validating CDDSS and maintaining their KBs, the future for CDDSS looks promising.

Acknowledgments Disclosure of (Non) Conflicts of Interest: Dr. Miller previously received royalties from the University of Pittsburgh in recognition of his work there in authoring the Quick Medical Reference (QMR)® program and knowledge base for diagnostic decision support in Internal Medicine; he donated most of those royalties to charity. The program is no longer being sold. Dr. Miller receives royalties from Vanderbilt University based on Vanderbilt's commercialization of the WizOrder CPOE system, and the Star electronic health record system, both of which he helped to develop. The majority of income from licensing these systems goes directly to Vanderbilt School of Medicine, per se. Dr. Miller does not own stock in, or serve as an employee or officer of, any of those vendors' companies. Work on INTERNIST-1 and QMR has been supported over the years by NIH grants from DRR and NLM; grants from Paul Mongerson and his CAMDAT Foundation; and, by internal funding from the University of Pittsburgh. In addition to the author, major contributors to the INTERNIST-1 and QMR projects include: Jack D. Myers, MD; Harry E. Pople, Jr., PhD; Griff Smith, PhD; Zachary Moraitis, MD; Craig Dean; Casey Quayle; Chuck Olson, PhD; Fred E. Masarie, Jr., MD; Nunzia B. Giuse, MD, MLS; Dario A. Giuse, Dr. Ing; Richard A. Bankowitz, MD; and numerous medical students and faculty members from the University of Pittsburgh and other institutions.

References

Adams, I. D., Chan, M., Clifford, P. C., et al. (1986). Computer aided diagnosis of acute abdominal pain: A multicentre study. *British Medical Journal (Clinical research ed.), 293*, 800–804.

Aliferis, C. F., Cooper, G. F., Miller, R. A., Buchanan, B. G., Bankowitz, R., & Giuse, N. B. (1996). A temporal analysis of QMR. *Journal of the American Medical Informatics Association, 3*, 79–91.

Bankowitz, R. A., McNeil, M. A., Challinor, S. M., Parker, R. C., Kapoor, W. N., & Miller, R. A. (1989a). A computer-assisted medical diagnostic consultation service: Implementation and prospective evaluation of a prototype. *Annals of Internal Medicine, 110*, 824–832.

Bankowitz, R. B., McNeil, M. A., Challinor, S. M., & Miller, R. A. (1989b). Effect of a computer-assisted general medicine diagnostic consultation service on housestaff diagnostic strategy. *Methods of Information in Medicine, 28*, 352–356.

Barnett, G. O., Cimino, J. J., Hupp, J. A., & Hoffer, E. P. (1987). DXplain. An evolving diagnostic decision-support system. *JAMA, 258*, 67–74.

Bates, D. W., et al. (2003). Ten Commandments for effective clinical decision support. *Journal of the American Medical Informatics Association, 10*(6), 523–530.

Berner, E. S., Webster, G. D., Shugerman, A. A., et al. (1994). Performance of four computer-based diagnostic systems. *New England Journal of Medicine, 330*(25), 1792–1796.

Berner, E. S., Maisiak, R. S., Cobbs, C. G., & Taunton, O. D. (1999). Effects of a decision support system on physicians' diagnostic performance. *Journal of the American Medical Informatics Association, 6*(5), 420–427.

Bleich, H. L. (1969). Computer evaluation of acid-base disorders. *Journal of Clinical Investigation, 48*, 1689–1696.

Blois, M. S., Tuttle, M. S., & Sherertz, D. D. (1981). RECONSIDER: A program for generating differential diagnoses. In: Hefferman, H. G. (Ed.). *Proceedings of the Fifth Annual Symposium On Computer Applications in Health Care* (pp. 263–268). Washington, DC: IEEE Computer Society Press.

Collen, M. F. (1995). *A history of medical informatics in the United States: 1950 to 1990. Bethesda* (p. 489). Amer Med Informatics Assoc: MD. ISBN 0-9647743-0-5.
Cooper, G. F. (1986). A diagnostic method that uses causal knowledge and linear programming in the application of Bayes' formula. *Computer Methods and Programs in Biomedicine, 22*(2), 223–237.
de Dombal, F. T., Hortocks, J. C., Staniland, J. R., & Gill, P. W. (1971). Simulation of clinical diagnosis: A comparative study. *British Medical Journal, 2*, 575–577.
Durack, D. T. (1978). The weight of medical knowledge. *New England Journal of Medicine, 298*(14), 773–775.
Feldman, M. J., & Bartlett, G. O. (1991). An approach to evaluating the accuracy of' DXplain. *Computer Methods and Programs in Biomedicine, 35*, 261–266.
Friedman, C. P., & Wyatt, J. (1997). *Evaluation methods in medical informatics* (p. 311). New York: Springer. ISBN 0387942289, 9780387942285.
Friedman, C. P., Elstein, A. S., Wolf, F. M., et al. (1999). Enhancement of clinicians' diagnostic reasoning by computer-based consultation: A multisite study of 2 systems. *JAMA, 282*(19), 1851–1856.
Fryback, D. G. (1978). Bayes' theorem and conditional nonindependence of data in medical diagnosis. *Computers and Biomedical Research, 11*, 423–434.
Giuse, N. B., Bankowitz, R. A., Giuse, D. A., Parker, R. C., & Miller, R. A. (1989a). Medical Knowledge Base Acquisition: The Role of Expert Review Process in Disease Profile Construction. *Proceedings of Thirteenth Annual Symposium on Computer Applications in Medical Care* (pp. 105–109). Washington, DC: IEEE Press.
Giuse, N. B., Giuse, D. A, & Miller, R. A. (1989b). Medical knowledge base construction as a means of introducing students to medical informatics. *Proceedings of the International Symposium on Medical Informatics and Education* (pp. 228–232). Victoria, BC.
Giuse, D. A., Giuse, N. B., & Miller, R. A. (1990). Towards computer assisted maintenance of medical knowledge bases. *Artificial Intelligence in Medicine, 2*, 21–33.
Giuse, N, B, Giuse, D. A, Bankowitz, R. A., & Miller, R. A. (1993a). Comparing contents of a knowledge base to traditional information sources. *Proceedings of the Seventeenth Annual Symposium on Computer Applications in Medical Care.* Washington DC: McGraw-Hill, Nov. 1993.
Giuse, D. A., Giuse, N. B., & Miller, R. A. (1993b). Consistency enforcement in medical knowledge base construction. *Artificial Intelligence in Medicine, 5*, 245–252.
Giuse, N. B., Giuse, D. A., Miller, R. A., Bankowitz, R. A., Janosky, J. E., Davidoff, F., et al. (1993c). Evaluating consensus among physicians in medical knowledge base construction. *Methods of Information in Medicine, 32*, 137–145.
Giuse, D. A., Giuse, N. B., & Miller, R. A. (1995). Evaluation of long-term maintenance of a large medical knowledge base. *Journal of the American Medical Informatics Association, 2*, 297–306.
Gorman, P. N., & Helfand, M. (1995). Information seeking in primary care: How physicians choose which clinical questions to pursue and which to leave unanswered. *Medical Decision Making, 15*(2), 113–119.
Gorry, A. (1968). Strategies for computer-aided diagnosis. *Mathematical Biosciences, 2*, 293–318.
Gorry, G. A., & Barnett, G. O. (1968). Experience with a model of sequential diagnosis. *Computers and Biomedical Research, 1*, 490–507.
Horrocks, J. C., McCann, A. P., Staniland, J. R., Leaper, D. J., & de Dombal, F. T. (1972). Computer-aided diagnosis: Description of an adaptable system, and operational experience with 2, 034 cases. *British Medical Journal, 2*(5804), 5–9.
Hupp, J. A., Cimino, J. J., Hoffer, E. F., Lowe, H. J., & Barnett, G. O. (1986). Explain-A computer-based diagnostic knowledge base. In: *Proc Fifth World Conference on Medical Informatics, MEDINFO 86* (pp. 3117–3121). Amsterdam: North-Holland.
Kingsland, L. C. III, Sharp, G. C., Kay, D. R., Weiss, S. M., Roeseler, G. C., & Lindberg, D. A. B. (1982). An expert consultant system in rheumatology: AI/RHEUM. *Proc Sixth Ann Symp Comput Appl Med Care* (pp. 748–752).
Kingsland, L., Sharp, G., & Capps, R. (1983). Testing of a criteria-based consultant system in rheumatology. In J. van Bemmel, M. Ball, O. Wigertz, et al. (Eds.), *Proceedings of MEDlNFO-83* (pp. 514–517). Amsterdam, The Netherlands: North-Holland.
Lau, L. M., & Warner, H. R. (1992). Performance of a diagnostic system (Iliad) as a tool for quality assurance. *Computers and Biomedical Research, 25*, 314–323.
Ledley, R. S., & Lusted, L. B. (1959). Reasoning foundations of medical diagnosis. *Science, 130*, 9–21.
Leigh, T. M., Young, P. R., & Haley, J. V. (1993). Performances of family practice diplomates on successive mandatory recertification examinations. *Academic Medicine, 68*(12), 912–918.
Lindbeg, D. A. B., Rowland, L. R., Buch, C. R. Jr., Morse, W. F., & Morse, S. S. (1968). CONSIDER: A computer program for medical instruction. *Proc Ninth IBM Med Symp.*

Lipkin, M., & Hardy, J. D. (1958). Mechanical correlation of data in differential diagnosis of hematological diseases. *JAMA, 166*, 113–123.

Madlon-Kay, D. J. (1989). The weight of medical knowledge: Still gaining. *New England Journal of Medicine, 321*(13), 908.

Masarie, F. E., Jr., Miller, R. A., & Myers, J. D. (1985). INTERNIST-I PROPERTIES: Representing common sense and good medical practice in a computerized medical knowledge base. *Computers and Biomedical Research, 18*, 458–479.

Miller, R. A. (1984). Internist-1/CADUCEUS: Problems facing expert consultant programs. *Methods of Information in Medicine, 23*, 9–14.

Miller, R. A. (1990). Why the standard view is standard: People, not machines, understand patients' problems. *Journal of Medicine and Philosophy, 15*, 581–591.

Miller, R. A. (1994). Medical diagnostic decision support systems past, present, and future. *Journal of the American Medical Informatics Association, 1*, 8–27.

Miller, R. A. (1996). Evaluating evaluations of medical diagnostic systems. *Journal of the American Medical Informatics Association, 3*, 429–431.

Miller, R. A. (1997). A heuristic approach to the multiple diagnoses problem. In E. Keravnou, C. Garbay, R. Baud, & J. Wyatt (Eds.), *Artificial intelligence in medicine, proceedings of AIME 97. Lecture notes in artificial intelligence series, #1211* (pp. 187–198). Berlin: Springer.

Miller, R. A., & Gardner, R. M. (1997). Summary recommendations for responsible monitoring and regulation of clinical software systems. *Annals of Internal Medicine, 127*(9), 842–845.

Miller, R. A., & Masarie, F. E., Jr. (1989). Use of the quick medical reference (QMR) (R) program as a tool for medical education. *Methods of Information in Medicine, 28*, 340–345.

Miller, R. A., & Masarie, F. E., Jr. (1990). The demise of the "greek oracle" model for medical diagnostic systems. *Methods of Information in Medicine, 29*, 1–2.

Miller, R. A., & Masarie, F. E. (1992). The Quick Medical Reference (QMR) Relationships function: Description and evaluation of a simple, efficient "multiple diagnoses" algorithm. *Proc MEDINFO 92* (pp. 512–518). Geneva, Switzerland.

Miller, R. A., & Schaffner, K. F. (1982). The logic of problem-solving in clinical diagnosis: A course for second-year medical students. *Journal of Medical Education, 57*, 63–65.

Miller, R. A., Pople, H. E., Jr., & Myers, J. D. (1982). INTERNIST-1, an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine, 307*, 468–476.

Miller, R. A., Schaffner, K. F., & Meisel, A. (1985). Ethical and legal issues related to the use of computer programs in clinical medicine. *Annals of Internal Medicine, 102*, 529–536.

Miller, R. A., Masarie, F. E., & Myers, J. D. (1986a). "Quick medical reference" for diagnostic assistance. *MD Computing, 3*, 34–48.

Miller, R. A., McNeil, M. A., Challinor, S., Masarie, F. E., & Myers, J. D. (1986b). Status report: The INTERNIST-1/quick medical reference project. *Western Journal of Medicine, 145*, 816–822.

Nash, F. A. (1954). Differential diagnosis: An apparatus to assist the logical faculties. *Lancet, 1*, 874.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.

Newman-Toker, D. E., & Pronovost, P. J. (2009). Diagnostic errors—The next frontier for patient safety. *JAMA, 301*(10), 1060–1062.

Nordyke, R. A., Kulikowski, C. A., & Kulikowski, C. W. (1971). A comparison of methods for the automated diagnosis of thyroid dysfunction. *Computers and Biomedical Research, 4*(4), 374–389.

Osheroff, J. A., Forsythe, D. E., Buchanan, B. G., Bankowitz, R. A., Blumenfeld, B. H., & Miller, R. A. (1991). Physicians' information needs: An analysis of questions posed during clinical teaching in internal medicine. *Annals of Internal Medicine, 114*, 576–581.

Parker, R. C., & Miller, R. A. (1989). Creation of a knowledge base adequate for simulating patient cases: Adding deep knowledge to the INTERNIST-1/QMR knowledge base. *Methods of Information in Medicine, 28*, 346–351.

Pauker, S. G., Gorry, G. A., Kassirer, J. P., & Schwartz, W. B. (1976). Towards the simulation of clinical cognition. Taking a present illness by computer. *American Journal of Medicine, 60*(7), 981–996.

Pearl, J. (1987). Evidential reasoning using stochastic simulation of causal models. *Artificial Intelligence, 32*, 245–252.

Pople, H. E., Jr. (1982). Heuristic methods for imposing structure on ill-structured problems: The structuring of medical diagnostics. In P. Szolovits (Ed.), *Artificial intelligence in medicine* (pp. 119–190). Boulder, Co: Westview Press. AAAS Symposium Series, no. 51.

Pople, H. E., Myers, J. D., & Miller, R. A (1975). DIALOG: A model of diagnostic logic for internal medicine. In: *Proceedings of the fourth International Joint Conference on Artificial Intelligence* (pp. 848–855). Cambridge, Massachusetts: MIT Artificial Intelligence Laboratory Publications.

Porter, J. F., Kingsland, L. C., I. I. I., Lindbeerg, D. A., et al. (1988). The AI/RHEUM knowledge-based computer consulatant system in rheumatology. Performance in the diagnosis of 59 connective tissue disease patients from Japan. *Arthritis and Rheumatism, 31*, 219–226.

Ramnarayan, P., Kapoor, R. R., Coren, M., Nanduri, V., Tomlinson, A. L., Taylor, P. M., et al. (2003). Measuring the impact of diagnostic decision support on the quality of clinical decision making: Development of a reliable and valid composite score. *Journal of the American Medical Informatics Association, 10*(6), 563–572.

Ramnarayan, P., Tomlinson, A., Kulkarni, G., Rao, A., & Britto, J. (2004). A novel diagnostic aid (ISABEL): Development and preliminary evaluation of clinical performance. *Studies in Health Technology and Informatics, 107*(Pt 2), 1091–1095.

Ramnarayan, P., Winrow, A., Coren, M., Nanduri, V., Buchdahl, R., Jacobs, B., et al. (2006). Diagnostic omission errors in acute paediatric practice: Impact of a reminder system on decision-making. *BMC Medical Informatics and Decision Making, 6*, 37–39.

Ramsey, P. G., Carline, J. D., Inui, T. S., Larson, Lo. Gerfo. J. P., Norcini, J. J., & Wenrich, M. D. (1991). Changes over time in the knowledge base of practicing internists. *JAMA, 266*(8), 1103–1107.

Shiffman, R. N. (1995). Guideline maintenance and revision. 50 years of the Jones criteria for diagnosis of rheumatic fever. *Archives of Pediatrics and Adolescent Medicine, 149*(7), 727–732.

Shortliffe, E. H. (1976). *Computcr-based medical consultations: MYCIN*. Artificial Intelligence Series. New York: Elsevier Computer Science Library.

Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D., & Levy, S. (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics, 21*(5), 631–643.

Stead, W. W., Haynes, R. B., Fuller, S., Friedman, C. P., et al. (1994). Designing medical informatics research and library–Resource projects to increase what is learned. *Journal of the American Medical Informatics Association, 1*(1), 28–33.

Szolovits, P., Patil, R. S., & Schwartz, W. B. (1988). Artificial intelligence in medical diagnosis. *Annals of Internal Medicine, 108*, 7.

Warner, H. R., Jr. (1989). Iliad: Moving medical decision-making into new frontiers. *Methods of Information in Medicine, 28*, 370–372.

Warner, H. R., Toronto, A. F., Veasey, L. G., & Stephenson, R. A. (1961). Mathematical approach to medical diagnosis. *JAMA, 177*, 75–81.

Warner, H. R, Haug, P., & Bouhaddou, O., et al. (1987). ILIAD as an expert consultant to teach differential diagnosis. In: *Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care* (pp. 371–376). New York: IEEE Computer Society Press.