Probabilistic Diagnosis Using a Reformulation of the INTERNIST-1/QMR Knowledge Base

II. Evaluation of Diagnostic Performance

Blackford Middleton, M.P.H., M.D., M.S.

Michael Shwe, M.S.

David Heckerman, Ph.D.

Max Henrion, Ph.D.

Eric Horvitz, Ph.D.

Harold Lehmann, M.D.

Gregory Cooper, M.D., Ph.D.*

Section on Medical Informatics, MSOB X215,

Stanford University, Stanford, CA 94305-5479

***Section of Medical Informatics, B50A Lothrop Hall, 190 Lothrop Street**

University of Pittsburgh, Pittsburgh, PA 15261

Abstract: We have developed a probabilistic reformulation of the Quick Medical Reference (QMR) system. In Part I of this two-part series, we described a two-level, multiply connected belief-network representation of the QMR knowledge base and a simulation algorithm to perform probabilistic inference on the reformulated knowledge base. In Part II of this series, we report on an evaluation of the probabilistic QMR, in which we compare the performance of QMR to that of our probabilistic system on cases abstracted from continuing medical education materials from Scientific American Medicine. In addition, we analyze empirically several components of the probabilistic model and simulation algorithm.

Keywords: Expert Systems, Computer-Aided Diagnosis, Probabilistic Inference, Belief Networks

1. Introduction

Quick Medical Reference (QMR®) is a microcomputer-based decision-support tool for diagnosis in internal medicine. We are developing a decision-theoretic version of QMR, which we call QMR-DT.[1] In Part I of this series, we discussed the reasons for developing a decision-theoretic QMR, the belief-network representation that we have used in the QMR-DT knowledge base (KB), and the algorithms that we implemented for inference on the QMR-DT KB. We have focused our research to date on the probabilistic component of the QMR-DT system.

Following the convention used in Part I, we refer to the QMR-DT KB and the assumptions that it includes as the *QMR-DT model*. These assumptions, which we examined in Part I, are marginal independence of diseases, conditional independence of findings given any hypothesis of diseases, causal independence of the influence of multiple diseases on a single finding, and binary-valued findings and diseases. Thus, we distinguish the QMR-DT model from the algorithms that we use for probabilistic inference on the QMR-DT model.

Briefly, to review our algorithms, we aim to compute the posterior marginal probability $P(d_i^+| F)$ for all diseases d_i given a set of findings F, where d_i^+ is the event that disease d_i is present. Note that $P(d_i^+| F)$ does not assume that only d_i is present, but rather allows for multiple diseases to be present in the patient. Calculation of $P(d_i^+ | F)$ using the QMR-DT two-level belief network is an NP-hard problem, however [1].

We implemented an algorithm that uses Bayes' rule under the added assumption that diseases are mutually exclusive. We call this algorithm *tabular Bayes'* rule (TB), to

---

® QMR is a registered trademark of the University of Pittsburgh.

1 We are currently using the INTERNIST-1 KB (circa 1986), rather than the more recent QMR KB. These two KBs are quite similar, to the extent that the methods in this paper can be applied to the latter KB as well. For simplicity, where the distinction between the INTERNIST-1 KB and QMR KB is inconsequential, we will refer to the INTERNIST-1 KB as the QMR KB.

reflect the fact that it is Bayes' rule under this assumption and the assumption that findings are conditionally independent given any disease. We also implemented a heuristic algorithm, called *iterative tabular Bayes'* algorithm (ITB), which applies TB several times to a set of findings. We call the set of diseases concluded by ITB as likely disease candidates for a particular set of findings the *heuristic-importance set*.

We implemented a stochastic simulation algorithm to approximate the posterior marginal probabilities of diseases $P(d_i^+ | F)$, where we do not make the assumption that at most one disease can be present in a patient. We will henceforth use the notation $\hat{P}(d_i^+ | F)$ to denote the estimates of the posterior marginal probabilities of disease that the simulation algorithm produces. This stochastic simulation algorithm, which we call S, uses the output of ITB as a heuristic to improve the convergence properties of the simulation in computing $\hat{P}(d_i^+ | F)$. That is, we focus the attention of the simulation initially on the diseases recommended by ITB. In addition, algorithm S uses a technique called *self–importance sampling* to refine this focus as the simulation progresses. We refer the reader to Part I, Section 3, for a detailed discussion of the inference algorithms that we implemented.

In this paper, we report an evaluation of the diagnostic performance of QMR-DT on one set of test cases. We focus first on relatively simple test cases to determine if the model and algorithms produce acceptable results. Specifically, in Section 2, we describe how we selected and constructed the diagnostic test cases and how we compared various inference algorithms. In Section 3, we report the results of the evaluation of diagnostic performance and sensitivity analysis of the QMR-DT model and algorithms. In Section 4, we provide a discussion of the experiments that we performed. A description of the notation used in this paper appears in the Appendix.

2. Experimental Design

To investigate the relative diagnostic accuracy of QMR and QMR-DT, we compared the performance of QMR with that of S on a set of test cases. In the remainder of Section 2, we describe the test cases used in the experiment, techniques used to compare the performance of QMR and S, and sensitivity analyses testing various components of the model and the simulation algorithm.

2.1 Test Cases

To evaluate QMR and S, we used cases abstracted from the Scientific American Medicine (SAM) Continuing Medical Education service [2]. The SAM cases provide physicians with a means of reviewing current clinical practice and of keeping abreast of new developments in internal medicine. Cases are created following specific guidelines by an expert in the subspecialty area containing the diagnosis in the case. The cases typically contain diagnoses consisting of a single disease, or occasionally two or three diseases. After an expert creates a case, it is reviewed by the SAM editorial staff for accuracy and for consistency with other SAM cases.

Subscribers to the SAM Continuing Medical Education service use cases in the following manner. In either computer-based or paper format, an introduction to a clinical scenario is presented. The reader is then given a variety of choices to obtain additional information. New information is not revealed unless the reader selects an item on the computer display or on paper by using a special marker. Depending on the items selected, the user follows one of several paths in a branching logic to make a diagnosis and to choose therapy. For didactic purposes, each information item has a number associated with it that represents an ad hoc rating of the clinical usefulness of the information for diagnostic or management purposes in the case. These ratings are made initially by the case author, and are reviewed by the editors of SAM. A score to assess performance in the case is created by summing the positive and negative weights associated with each information item selected.

In abstracting the SAM cases, we selected only those findings that had a positive score—that is, those findings whose presence or absence was relevant to diagnosis or management in the case. We also limited the abstracted findings to those that typically would be available after initial evaluation and workup in a physician's office, emergency room, or outpatient clinic. We attempted to simulate the diagnostic dilemma faced by a doctor in the early stages of the clinical course of disease. Thus, we did not include information that would be available only after an extended diagnostic workup. Typically, this set of information comprises a fraction of the available information in a SAM case. Much of the information in the SAM cases pertains to evolving patient management, a large part of the SAM educational exercise. The information pertaining to patient management, however, is not required by a clinician or diagnostic system to make the diagnoses that we listed in the reference diagnoses for a SAM case.

For the purpose of this study, we define the *reference diagnosis* of a SAM case to be those diseases specified by SAM as the patient's diagnosis, such that the findings for each of the diseases in the diagnosis were present on initial evaluation of the patient. Additional diagnoses may be specified by SAM as occurring in a case sometime later in the clinical course. For instance, if other findings appeared after the initial evaluation, leading to an additional diagnosis, we did not include the additional diagnosis in the reference diagnosis for the case.

The SAM case findings were first translated into the vocabulary of QMR for analysis of QMR and then into INTERNIST-1 terms, if possible, for analysis of QMR-DT. Recall the distinction between the two knowledge bases: INTERNIST-1 was used to create QMR-DT. QMR was used to evaluate the performance of QMR-DT. Because QMR is the successor of INTERNIST-1, it contains a more contemporaneous vocabulary for describing findings, and additional findings. Although the abstraction process was generally straightforward, we encountered three types of difficulties: (1) mapping a negative (absent or normal) finding to INTERNIST-1 and QMR terms when the finding occurs only as a positive finding in the INTERNIST-1 KB and QMR KB; (2) mapping broad, categorical findings; and (3) mapping findings for which there does exist a QMR label, but does not exist a sufficiently congruent INTERNIST-1 label. We discuss each difficulty in turn.

Most INTERNIST-1 and QMR findings do not contain specific categories to denote a finding as normal. For instance, to denote an abnormal level of serum cholesterol, we may choose from either of the two binary findings: *CHOLESTEROL BLOOD INCREASED* or *CHOLESTEROL BLOOD DECREASED*. But to specify that the cholesterol value is *normal,* we must specify that both of the abnormal cholesterol findings are negative [3]. Next, consider findings corresponding to continuous variables for which there exist three or more INTERNIST-1/QMR descriptors. For example, to specify an abnormal serum glutamic oxaloacetic transaminase (SGOT) level, we may choose from any of the three findings: *SGOT* 40 *TO* 119, *SGOT* 120 *TO* 400, or *SGOT GTR THAN* 400. To specify a normal SGOT value, we record *SGOT* 40 *TO* 119 as a negative finding; that is, we indicate as negative the least abnormal choice of the finding of interest to represent a normal finding. This procedure is acceptable when, in the normal course of a disease process, all levels of abnormality may be expected; thus, if any one abnormal level is present, the disease should be considered as a possible diagnosis**.** Problems may arise, however, when not all values may be expected in the course of the disease, or when an extreme abnormal finding is required to cause a disease to be considered or rejected. For example, consider a disease profile containing the finding *SGOT GTR THAN* 400 but not *SGOT* 120 *TO* 400, or *SGOT GTR THAN* 400. Diseases of this type may not be penalized as diagnostic possibilities because we enter only the least abnormal finding as negative or absent. When a negative finding f_j is entered, the posterior probability of a disease d_i is diminished only if f_j is in the profile of d_i . We expect this problem to be small, because most disease profiles in the

INTERNIST-1 KB have all categories of multivalued findings entered on relevant disease profiles [4]. For example, the disease profile of HEPATITIS ACUTE VIRAL includes the three findings *SGOT* 40 *TO* 119, *SGOT* 120 *TO* 400, or *SGOT GTR THAN* 400 which together encompass the entire range of abnormal SGOT values.

SAM cases often report negative categorical findings that represent broad concepts in the history and physical examination. However, there is no convenient way in QMR or INTERNIST-1 to represent, for example, a negative review of systems. When specific findings from a review of systems were relevant to a SAM case, the findings were typically offered in the case as information items that could be selected by the reader. If the case gave only a negative review of systems without subsequently detailing findings relevant to the case, we did not use this information. Also, we did not infer negative findings unless they appeared as specific information items in the case.

Next, consider the QMR finding *BRAIN CT SELLA TURCICA ENLARGED*. This finding is an example of simple lack of congruence between a QMR finding and findings in INTERNIST-1, since it appears in the QMR finding hierarchy but is absent from the INTERNIST-1 hierarchy. The INTERNIST-1 KB contains the finding *SKULL XRAY SELLA ENLARGED*, which we used as a surrogate for the finding *BRAIN CT SELLA TURCICA ENLARGED.*

A more complicated example of lack of congruence between QMR and INTERNIST-1 findings is a finding about the absence of heart murmur. QMR represents the presence of heart murmur with a specific finding, *HEART MURMUR PRESENT*. This term allows us to indicate the absence of heart murmur by recording this finding as negative in QMR. By contrast, INTERNIST-1 does not have the high-level representation *HEART MURMUR PRESENT*. It is possible to indicate that all different types of heart murmur are negative but to do so we would have to enter more than 40 different types of heart murmurs as negative. Thus, to represent the absence of heart murmur using INTERNIST-1 terminology, we record *HEART MURMUR SYSTOLIC EJECTION LEFT STERNAL BORDER* as a negative finding. In this and other similar situations we attempted to select a common INTERNIST-1 finding that was associated with many of the same diseases associated with the QMR finding. Some important diseases, however, that do not contain the surrogate finding in their disease profiles will not be affected by the presence or absence of the surrogate finding. In those cases where there did not exist an INTERNIST-1 finding that was identical to the QMR finding, we ran QMR with the best possible mapping of case findings into QMR terminology, and ran the QMR-DT algorithms using the closest mapping into INTERNIST-1 terms.

In this evaluation of diagnostic performance we analyzed only test cases containing a single disease in the reference diagnosis. We sought to determine if the QMR-DT model and simulation algorithm gave adequate results on relatively straightforward cases. In a subsequent study we plan to evaluate the performance of QMR-DT on more complex, multiple-disease cases. The static nature of our test cases makes the usual hypothetico-deductive approach to diagnosis impossible. A physician or a computer–based diagnostic aid cannot iteratively hypothesize and refine a differential diagnosis with new evidence when all of the case evidence is given at once. QMR is intended to be used in a "mixed–initiative" manner between system and physician user [5]. Unlike INTERNIST-1, which has a partitioning algorithm that allows the system to focus iteratively on different problem areas [6], QMR applies a scoring scheme once, and then provides other options to the user for solving difficult cases [7]. Because we could not *a priori* define an experimental protocol to take advantage of these options in a controlled fashion, we were unable to employ a mixed–initiative approach in our evaluation of QMR performance. Thus, QMR performance may not reach its optimal level of accuracy in our evaluation. We believe, nevertheless, that similar limitations apply to the QMR-DT test algorithms and that the comparison is useful.

Of the total of 62 SAM cases that were made available to us, we rejected 15 because the reference diagnosis was not contained in the QMR KB, and six others because the primary diagnosis did not appear in the INTERNIST-1 KB. Of the remaining 41 SAM cases, 38 contained single-disease diagnoses. We rejected three SAM cases with multiple disease diagnoses because of the difficulty of properly analyzing these cases with a mixed–initiative or hypothetico-deductive approach. We used 15 of the 38 remaining cases to test the various inference algorithms while we were developing them. We reserved 23 of the 38 cases for our evaluation study. None of the cases in the set of 23 was presented to any of the algorithms before the final evaluation. Information on the diagnoses and findings of the 23 SAM cases appears in Table 1.

We did not randomly sample from the set of 38 cases to create the set of evaluation cases. Rather, we used for development and testing the first 15 applicable SAM cases that we received from Scientific American Medicine, and we used for evaluation the cases that we received after we began testing.

2.2 Comparison of Rank Ordering

We are interested primarily in evaluating the diagnostic accuracy of QMR and QMR-DT. By *diagnostic accuracy*, we mean how high an algorithm ranks the reference diagnoses on a differential diagnosis. According to this definition, an algorithm would be perfectly accurate if it ranked the reference diagnosis highest in all cases. We limit our comparison in this study to ranks because QMR does not produce a probabilistic differential. The version of QMR we used produces two different types of diagnostic opinion: (1) a set of "potentially interesting diagnostic hypotheses," which consists of a rank-ordered list of diseases and QMR scores; and (2) one or more "unifying hypotheses," each of which consists of a primary diagnosis with various possible antecedent and consequent diseases. In this paper, we limit our analysis of QMR's performance to the rank-ordered list of diseases.

To compare diagnostic algorithms, we used a two-sided Wilcoxon signed-rank test [8] in pairwise comparisons of rank ordering of diseases. Specifically we used the Wilcoxon test to compare the rank orderings of algorithm S with those of QMR and ITB. For each pairwise comparison, the null hypothesis is that the rank orderings of the algorithms are the same.

2.3 Sensitivity Analyses

We distinguish a sensitivity analysis on the *QMR-DT model* from a sensitivity analysis on the *simulation algorithm*. We refer to a sensitivity analysis on the model as one in which we vary either the connectivity or probabilities of the belief network that we call the QMR-DT model, to produce a second model, QMR-DT$'$. Thus, given a set of findings F, we would not expect the posterior distribution implied by the QMR-DT$'$ model to be equal to the posterior distribution implied by the QMR-DT model. By contrast, in a sensitivity analysis of the simulation algorithm, we hold the QMR-DT model constant and change one of the components of the algorithm. Accordingly, with the admissible simulation algorithms we shall apply we expect that, in the limit, the estimates $\hat{P}(d_i+ \mid F)$, for each disease d_i, will converge to the values $P(d_i+ \mid F)$ implied by the QMR-DT model. However, after a finite amount of simulation, $\hat{P}(d_i+ \mid F)$ may deviate significantly from $P(d_i+ \mid F)$.

2.3.1 Analysis of the QMR-DT Model

In our sensitivity analysis of the QMR-DT model, we present the effect of each of the following assumptions on the differential output of the S algorithm: uniform leak probabilities, uniform prior probabilities of disease, and mutually exclusive diseases. We discuss each of these analyses in turn.

Recall from Part I, Section 2.2.3 that the leak probability represents the probability that a finding is caused either spontaneously (e.g., a false positive) or by a disease not modeled in the QMR KB. We performed the analysis of uniform leak probabilities to investigate the influence on performance provided by the leak probabilities that we derived. We shall use the term *S/UL* to refer to S running on the model QMR-DT$'$ where $P(f^+ \mid \text{only } L_f) = 10^{-5}$, and L_f is the leak event. The lowest prior probability of diseases in the QMR KB is approximately 2×10^{-5}. We used the value of 1×10^{-5} because the leak probabilities generally should be lower than the prior probability of diseases that are modeled in the QMR KB to avoid over representation of the leak event.

We perform the analysis of uniform disease prior probabilities to investigate whether the prior probabilities in the QMR-DT model derived from the National Center for Health Statistics hospital discharge data enhanced the performance of the system. To examine the diagnostic behavior of the QMR-DT model under the added assumption of uniform prior probabilities of diseases, we apply the S algorithm using uniform prior probabilities of diseases to estimate $P(d_i^+ \mid F)$. We shall use the term *S/UD* to refer to S running on a model QMR-DT$'$, where each disease is assigned a prior probability of $P(d_i^+)=10^{-3}$. There are many other values that we could have used as the uniform prior probability of disease. However, this value would not change the rank order in the values of $P(d_i^+ \mid F)$. Note that the leak probabilities for S/UD also were calculated using uniform prior probabilities on diseases.

In addition, we examined the performance of the QMR-DT model under the assumption of mutually exclusive diseases. Under this assumption, we can use TB to compute $P(\text{only } d_i^+ \mid F, \mu)$, where μ is the assumption of mutually exclusive diseases. We performed this analysis as a point of comparison to investigate the influence provided by modeling the possibility that a patient may have more than one disease. Recall the equation for computing $P(\text{only } d_i^+ \mid F, \mu)$:

$$P(\text{only } d_i^+ \mid F, \mu) = \frac{P(F \mid \text{only } d_i^+)\, P(\text{only } d_i^+)}{\sum_{k=1}^{n} P(F \mid \text{only } d_k^+)\, P(\text{only } d_k^+)} \tag{4}$$

Note that Equation 4 includes a term for the prior probability of a single disease, $P(\text{only } d_k^+)$. Since the NCHS data that we are using allows us to compute only prior probabilities of the form $P(d_k^+)$, which allow other diseases to be present in a patient, we use $P(d_k^+)$ as a proxy for $P(\text{only } d_k^+)$.

To test statistically the sensitivity analyses, we compare, using the Wilcoxon signed-rank test, the rank assigned to the reference diagnosis in a case by S/UL, S/UD, and TB to that assigned by S. We test the null hypothesis that the rank orderings (of the reference diagnosis) produced by two algorithms are identical.

2.3.2 Analysis of the Algorithm S

In a sensitivity analysis of algorithm S, we compare the probabilistic output of two modified versions of S to the probabilistic output of S itself. Thus, we use the posterior distribution of S as our reference distribution. Recall from Part I, Section 3.3 that S uses both heuristic priming from ITB and self-importance sampling. We refer to the S algorithm with no self-importance sampling as *S/NSI*. Similarly, we refer to the S algorithm with no ITB heuristic as *S/NITB*. Like S, S/NSI obtains its initial importance distribution P_0' from the heuristic-importance set of ITB. By contrast, S/NITB does not use the heuristic-importance set to generate P_0'. Rather, S/NITB sets $P_0'(d_j^+)$ for all d_j to the greater of 10^{-3} or the prior probability on d_j. S/NITB uses the same self-importance updating function as does S.

In addition to comparing S to S/NSI and S/NITB, we compared S to a second run of S, which we will call *S2*. We compared the probability distributions generated by S and S2 to examine the reproducibility of the simulation estimates. Note that reproducibility of the posterior distributions using the same simulation algorithm (with a different random number seed) is a necessary but insufficient condition for proof of convergence of the estimates to the posterior distribution of QMR-DT. Since S and S2 were both run for the same number of trials per case, we arbitrarily select S as the reference algorithm.

To compare the posterior distributions of S to S/NSI, S/NITB, and S2, we use a measure of the correlation of the two distributions over the ten diseases that S determines to have the highest posterior marginal probabilities. Let $d_{A(\mathrm{i})}$ be the disease assigned the *i*th rank by algorithm *A*. Thus, for example, $d_{A(1)}$ is the most probable disease according to algorithm *A*. Let $P_X(d^+_{A(i)} \mid F)$ be the probability that algorithm *X* assigns to disease $d_{A(i)}$ given the finding set *F*. We define the correlation $r(A, B)$ as the correlation coefficient over the pairs $(P_A(d^+_{A(i)} \mid F)\ , P_B(d^+_{A(i)} \mid F))$ for $1 \leq i \leq 10$. For example, to compare the posterior marginal probabilities generated by algorithm S to those generated by algorithm S2 on a specific test case, we compare the correlation coefficient over the pairs $(P_{\mathrm{S}}(d^+_{\mathrm{S}(i)} \mid F), P_{\mathrm{S2}}(d^+_{\mathrm{S}(i)} \mid F))$ for $1 \leq i \leq 10$. In general, $r(A, B)$ is not symmetric.

Because of the large number of diseases in the QMR-DT KB, when we run the diagnostic algorithms on the SAM cases, we record only the posterior marginal probabilities of the diseases ranked in the top 20 positions by any particular algorithm. If the rank assigned by algorithm *B* is greater than 20 for any $d_{A(\mathrm{i})}$, such that $1 \leq i \leq 10$, then we bound $P_B(d^+_{A(i)} \mid F)$ between 0 and $P_B(d^+_{A(20)} \mid F)$. In such cases, we use $P_B(d^+_{A(20)} \mid F)\ /\ 2$ as the value for $P_B(d^+_{A(i)} \mid F)$, since $P_B(d^+_{A(20)} \mid F) / 2$ is the expected value of $P_B(d^+_{A(i)} \mid F)$, assuming that $P_B(d^+_{A(i)} \mid F)$ is symmetrically distributed between 0 and $P_B(d^+_{A(20)} \mid F)$. For example, suppose that the *j*th-ranked disease of S does not appear in the top 20 ranked diseases of S2, and that $P_{\mathrm{S2}}(d^+_{\mathrm{S}(20)} \mid F) = 0.01$. Then, we use the value of 0.005 for $P_{\mathrm{S2}}(d^+_{\mathrm{S}(j)} \mid F)$.

We used a matched-pair *t* test to examine the difference in two correlation coefficients from two algorithms [8]. Using the SAM cases, we test two null hypotheses with a two-tailed matched-pair *t* test with a level of significance of $p = 0.05$. The first null hypothesis is that the absence of the self-importance sampling from S does not degrade significantly the performance of the algorithm. We test this hypothesis with a matched-pair *t* test to investigate whether the correlations of posterior probabilities between S/NSI versus S is equal to those between S2 versus S. The second null hypothesis is that the absence of the heuristic-importance set generated by ITB does not degrade significantly the performance of the S algorithm. We test this hypothesis by using the matched-pair *t* test to investigate whether the correlations of posterior probabilities between S/NITB versus S is equal to those between S2 versus S.

3. Results

We implemented TB, ITB, S, S/UD, S/UL, S/NITB, and S/NSI in LightSpeed Pascal on a Macintosh IIci. We used Version 10.729 of QMR with a version of the QMR KB that is dated 6/14/89. QMR running on a PS/2 Model 50 performed inference on each of the SAM cases in 5 to 20 seconds. For all the test cases that we ran for this study, TB required an average of 3 seconds (range 0.5 to 14 seconds) on each case, ITB required an average of 29 seconds per case (range 5 to 68 seconds) and S completed a total of 40,000 trials in an average of 94 minutes (range 46 to 173 minutes). (Note that the running times for S/UD, S/UL, S/NITB, S/NSI are similar to the running time of S.)

3.1. A Comparison of Ranks and an Analysis of the QMR-DT Model

After running these algorithms on the SAM cases, we record for each algorithm the ranks that the algorithm assigns to the diseases in the reference diagnosis of each test case. These ranks appear in Table 2; a "—" appears where an algorithm did not assign a rank to a disease in the reference diagnosis. We emphasize that QMR was not used in an interactive fashion and that the results presented herein represent its performance under constrained evaluation conditions [5]. For example, note that in test case 23 the reference diagnosis is "celiac sprue" (Table 1). When this case was analyzed with QMR it did not provide a rank for the reference diagnosis, however, it listed "malabsorption" as its topmost diagnosis (Table 2). In this instance QMR suggested a more general diagnosis, or intermediate pathophysiologic state, above the rank of the reference diagnosis. In this case, the diagnosis suggested by QMR was found to accurately represent the essential clinical state of a case yet it did not suggest the reference diagnosis. For the purposes of our study, however, we use only the ranks of the reference diagnosis and acknowledge that our results represent only an initial laboratory evaluation of the various models and algorithms [9].

We also summarize the ranking performance of each algorithm by noting the number of reference diagnoses ranked in the top position, the top five positions, the top 10 positions, and the top 20 positions. These summaries of the ranks appear in Table 3. The results of the Wilcoxon signed-rank test on the rank-ordering performance of the algorithms relative to S are shown in Table 4.

As discussed in Section 2.1, none of the 23 SAM cases in Table 1 was used during the development of the simulation algorithms. That is, the results reported in Table 2 were those obtained the first time that QMR, TB, ITB, S, S/UD, and S/UL were run on any of the SAM cases.[2]

---

[2] When we first ran SAM 45, we found that TB, ITB, S, and S/UL assigned to the reference diagnosis of primary aldosteronism a posterior probability of 0. The reason for this behavior is that the gender-adjusted prior probability assigned to primary aldosteronism was 0. Recall our assumption in Part I for assigning a prior probability to an age- or gender-specific category for which the NCHS statistics indicated that there were a negligible number of hospital discharges for a particular disease. The prior probability was calculated based on the discharges remaining after the discharges from other categories were subtracted from the general discharges. In the case of primary aldosteronism, the total number of patients discharged was listed as 3000, whereas the number of females discharged was 3000. Thus, our system inferred that no males were discharged with the diagnosis of primary aldosteronism and set to zero the prior probability of primary aldosteronism given that the patient is male. Only S/UD used a nonzero prior probability for this event, since it set all prior probabilities to 10^{-3}. When we discovered this error, we assigned the value of 1000 to the number of males discharged with primary aldosteronism. We then re-ran TB, ITB, S, and S/UL, S/NITB, and S/NSI on SAM 45. The lowest discharge value reported in the NCHS data was 2000. Thus, 1000 is the expected value of males discharged with primary aldosteronism, assuming a symmetric distribution of this value between 0 and 2000.

The primary aim of the evaluation in this paper is to compare the performance of S with QMR on the SAM cases. The summary of the ranks (Table 3) assigned by the two algorithms indicates that S performs comparably to QMR on the SAM cases. As shown in Table 4, the Wilcoxon signed-rank test failed at a level of statistical significance of $p = 0.05$ to reject the null hypothesis that the rank orderings of the two algorithms are identical on the SAM cases.

We also use the results of the rank-ordering performance on SAM cases in our sensitivity analysis on various components of the QMR-DT model. Let us first examine the difference in performance that we observe when we add the assumption of mutually exclusive disease hypotheses to the QMR-DT model. When we compare the performance summary of S to TB, we see that TB performs slightly better than S on the SAM cases (Table 3). The difference in performance, however, is not significant at the $p = 0.05$ level (Table 4).

We would expect TB to rank order the reference diagnoses at least as well as S on test cases with a reference diagnosis consisting of a single disease because TB assumes that at most one disease can exist in the patient. In other words, since the restrictive assumption of TB is compatible with the SAM test cases, the algorithm is tailored to the diagnostic task. Since each of the diagnoses of the cases in the SAM set contains a single disease, we introduce only additional degrees of freedom into the diagnostic algorithm by modeling the interaction of multiple diseases, as in algorithm S.

The rank-ordering data suggest either that the QMR-DT model is not sensitive to prior probabilities on diseases for the cases tested, or that the prior probabilities that we have assigned to the diseases are inaccurate. We see from Table 3 that S performed only slightly better than did S/UD on the SAM cases. The differences were not significant at the $p = 0.05$ level. However, in Table 4, we see in the sensitivity analysis of the leak probabilities a statistically significant difference ($p = 0.05$) on the SAM cases between the rank-ordering performance of S and S/UL. Note that on SAM cases 29, 30, 34, and 40 in

Table 2, S ranked the reference diagnosis in the top 20 of its differential, whereas S/UL placed the reference diagnoses much lower in its differential.

3.2 Analysis of the Simulation Algorithm S

In a sensitivity analysis of the simulation algorithm to its component heuristics, we compared the posterior distribution generated by S with the posterior distributions of S2, S/NSI, and S/NITB. Tables 5 shows the values for the correlation coefficient $r(A,B)$, as defined in Section 2.3. Each of the correlation coefficients $r(A,B)$ corresponds to a scatterplot of the data points $(P_A(d^+{}_{A(i)} \mid F)\ , P_B(d^+{}_{A(i)} \mid F))$ for $1 \leq i \leq 10$. To summarize graphically the similarity of a posterior distributions from two algorithms A and B, we overlay the scatterplots from several test cases to form an *aggregate scatterplot.* In Figure 1 appears the aggregate scatterplot for r(S, S2) on the SAM cases. Similarly, the aggregate scatterplot for r(S, S/NSI) appears in Figure 2, and the plot for r(S, S/NITB) appears in Figure 3.

Note that the posterior distributions of S and S2 were very similar over the top 10 diseases of the posterior distribution generated by S. We see from Figure 1 that the points in the aggregate scatterplot of S2 versus S lie close to the identity line. The similarity of the distributions of S and S2, as shown in Figure 1, supports the hypothesis that S is converging to the posterior distribution implied by QMR-DT.

To analyze the sensitivity of S to the ITB heuristic on the SAM cases, we compared the aggregate scatterplot of S/NITB versus S found in Figure 3 (pooled $r = 0.89$) to that of S2 versus S found in Figure 1 (pooled $r = 0.93$). The pooled r values indicate that S/NITB correlates with the distribution of S nearly as well as S2 does. The pairs of correlations r(S, S2) and r(S, S/NITB) for each of the SAM cases are not found to be significantly different by the two-tailed matched-pair t test ($p = 0.05$). This result suggests that absence of the ITB heuristic does not degrade significantly the convergence of simulation in the SAM cases. By contrast, when we compare the aggregate scatterplot in Figure 2 of S/NSI versus S (pooled $r = 0.81$) to the plot in Figure 1 of S2 versus S (pooled $r = 0.93$), we see that the absence of the self-importance heuristic on the SAM cases led to significant disparity between the estimates of S/NSI and S. The two-tailed matched-pair t test ($p = 0.05$) of the pairs of correlations r(S, S2) and r(S, S/NSI) for each of the SAM cases rejected the null hypothesis that the pairs of correlations were identical; this result suggests that the absence of the self-importance heuristic degrades the correlation with S. If we believe that the estimates of S are close to the posterior distribution implied by the QMR-DT model (as suggested by the similarity of the distributions of S and S2), then it seems that the estimates of S/NSI have not converged to the posterior distribution implied by the QMR-DT model in SAM cases.

In summary, the S algorithm exhibited diagnostic accuracy that is comparable to that of QMR on the SAM cases. The difference in the rank-ordering performance of the two algorithms was not statistically significant at the $p = 0.05$ level using a Wilcoxon signed-rank test. We have evidence that the estimates of S have converged to the posterior distribution implied by the QMR-DT model, since the estimates of S2 are very similar to those of S. In our sensitivity analysis of the QMR-DT model, we observed that neither the added assumption of mutually exclusive diseases nor uniform prior probabilities on diseases caused a statistically significant difference in the rank ordering of the reference diagnosis of the SAM cases. The assumption of uniform leak probabilities did, however, degrade the performance of S in the SAM cases to a significant ($p = 0.05$) extent. In our sensitivity analysis of the two heuristics used by S, we observed that the estimates of S/NSI (S without self-importance updating) were markedly different from those of S. Estimates of S/NITB (S without the heuristic iterative tabular Bayes' algorithm) were not significantly different from those of S.

4. Discussion

In this study, we reformulated the QMR KB into a probabilistic model (the QMR-DT KB) using a belief-network representation. We compared the performance of QMR to an implementation of stochastic simulation on the QMR-DT model. We found that, on single-diagnosis cases abstracted from continuing medical education materials from Scientific American Medicine, the simulation algorithm S performed comparably to the QMR diagnostic algorithm. In our sensitivity analysis of three components of the model, we found that only the assumption of uniform leak probabilities on findings resulted in a significant degradation in performance. In our analysis of the heuristics used by the S algorithm, we found that the algorithm was sensitive to the absence of a self-importance updating heuristic.

Noteworthy issues in this study include (1) the insensitivity of the QMR-DT model to uniform prior probabilities of diseases, (2) the value we used for the uniform leak probability, (3) the difficulties in abstracting cases, (4) the lack of a gold-standard distribution against which to compare our simulation algorithms, (5) the running time of the simulation, (6) the limitations of our rank-ordering comparison metric, (7) the limited use of QMR's diagnostic capabilities, and (8) the shortcomings of the QMR-DT belief network model. We address each of these concerns in turn.

4.1 Insensitivity to Prior Probabilities

We found that the QMR-DT model was insensitive to uniform prior probabilities of diseases. This result may be caused by noise in the prior probabilities of diseases that we used; our mapping between ICD-9-CM terms and INTERNIST-1 diseases was inexact. Alternatively, the prior probabilities of the diseases in the SAM cases may differ from those in the population from which our discharge statistics were collected. Also, it is possible that the weight of the evidence in the diagnostic cases that we used dominates the prior probabilities of diseases in deriving posterior probabilities of diseases.

4.2 Uniform Leak Probabilities

Recall from Part I, Section 2.2.3 that the leak probability of a finding is the probability that the finding occurs in the absence of any disease in the QMR KB, where L_f is the event that the leak is active. In our study of the effect of uniform leak probabilities, we used 10^{-5} as the value of $P(f^+ \mid L_f \text{ only})$ for each $f \in F^+$. We could have used other values for the uniform leak probability, such as the mean leak probability (1.8×10^{-3}) or the median leak probability (1.9×10^{-4}). Values of $P(f^+ \mid L_f \text{ only})$ other than 10^{-5} might have led to different rank orderings in the posterior marginal probabilities of diseases $P(d_i^+ \mid F)$ in the analysis of uniform leak probabilities. In any case, the observed effect of the uniform leak probabilities relative to the effect of the uniform prior probabilities on the diagnostic performance of the system suggests that QMR-DT model may be more sensitive to leak probabilities than to prior probabilities of diseases for the types of cases tested.

4.3 Abstracting Test Cases

We chose to use the SAM continuing medical education cases because they are in a standardized format that was amenable to abstraction, and because they have undergone review both by experts in the domain and by experts in test–case construction. These cases, however, are only an approximation of real clinical cases. The case creator or the case abstracter may introduce bias into the evidence set, due to differences in experience with similar clinical cases [10]. As described in Section 2.1, it was not always possible to map findings from a case into the finding representations present in QMR or INTERNIST-1. Because QMR's vocabulary is richer than that of INTERNIST-1, we encountered a number of findings that we could map into QMR terminology but not into INTERNIST-1 terminology. Thus, because our current QMR-DT belief network is based on the INTERNIST-1 KB, on occasion we had to use proxy findings from the INTERNIST-1 KB. Instead of running QMR on only the case findings that were translatable into INTERNIST-1 terminology, we ran QMR with the best possible mapping of case findings into QMR terminology. We attempted to reduce bias in case abstraction by following a standardized protocol, and by having a single reviewer for each case.

4.4 Lack of a Gold-Standard Posterior Distribution

Ideally, to examine the convergence properties of the simulation that we are using, we would like to know the posterior distribution implied by the QMR-DT model—that is, a gold-standard distribution. It is possible that Henrion's TopN algorithm [11] would be able to produce tight bounds on the posterior probabilities of diseases for large cases. In addition, the recursive decomposition algorithm [12] appears promising as an exact method to calculate the posterior probabilities of diseases in an acceptable amount of time in some cases. As these methods were under development at the time this project was initiated we chose to use a simulation algorithm. We are currently investigating these other possibilities.

One of the problems of the simulation algorithm that we describe in this paper is that the algorithm does not provide us with a measure of the error of its estimates of the posterior probabilities of diseases. In the absence of a gold-standard distribution and the absence of error measurements, we gain confidence in the estimates of the simulation when separate executions of the simulation on a specific set of evidence produce similar distributions. Although we report the results for only two runs of the simulation algorithm (S and S2), the close agreement between the distributions produced by S and S2 for each of the SAM cases gives us reason to believe that the estimates of S have converged to the posterior distribution implied by the QMR-DT model.

We also have used the S algorithm for inference on cases abstracted from clinicopathologic conference (CPC) cases. These cases may contain a large number of findings and multiple, co-existing diseases in their diagnoses. On many of these cases, we did not observe a high degree of correlation between the distributions of S and of S2. We continue to study the behavior of the simulation algorithm and work to improve the algorithm's convergence properties. One improvement that we have explored is Markov blanket scoring (MBS) [13, 14]. Our results indicate that the MBS modification increases the rate of convergence as a function of the number of trials [15]. In fact, we observed that separate executions of S with the MBS modification are able to reproduce the posterior probability distributions of disease in complex CPC cases.

4.5 Running Time of the Simulation Algorithm

We do not believe that the prolonged running time of our serial implementation of S on a personal computer will be a long-term limitation for practical applications. The simulation that we are using is readily amenable to parallelization. For example, we can decompose the simulation by trials (instantiations of the belief network), since each of the trials within one self-importance sampling interval is independent of the other trials. The running time of the simulation should decrease as a linear function of the number of processors. For example, a shared-memory parallel-computing machine with 64 processors, each with the computing power of a 68030 (the microprocessor in a Macintosh IIci), would decrease the time of computation by a factor of approximately 64, from approximately 94 minutes (the average running time of S on a Macintosh IIci on the cases reported in this study) to 1.5 minutes. Similar machines are currently accessible on the Internet for use in research such as the QMR-DT project.

Moreover, the simulation can incorporate any heuristic information to improve convergence time. (Conversely, a bad heuristic may degrade the convergence.) The S algorithm uses two heuristics: a set of approximately 20 diseases recommended by ITB as initial likely diagnostic candidates, and a self-importance sampling heuristic to update the sampling distribution based on the algorithm's current estimates of the posterior probabilities of diseases. In addition, we could incorporate information obtained from a physician familiar with a diagnostic case. For example, the physician could suggest diseases that he believes are likely in a patient given the findings observed. If the QMR-DT KB included a hierarchy of diseases (based on organ systems, for example), then the physician could also suggest the classes of the diseases that he believes to be present in the patient. In addition, he could provide the system with an estimate of the number of diseases that he believes to be present. Any of this information, if reasonably accurate, probably would improve the convergence time of the simulation.

4.6 Rank-ordering Evaluation Metric

In our evaluation of QMR-DT, we used as an evaluation metric the rank assigned by S or by QMR to the reference diagnosis. We realize that this metric is limited. The evaluation reported in this paper is part of an iterative cycle of test and refinement. We do not intend this analysis to be a definitive evaluation of QMR-DT or QMR. In a future more extensive comparison, we might want to use metrics that involve clinical opinion, or eventually patient outcome measures—when the systems are used clinically. For example, we might score the output of the two systems using the judgment of an expert regarding which differential most accurately reflects the probable state of the patient in light of the findings presented. Also, we could examine the effect of the system's differential on a physician's diagnostic beliefs or work-up plans. Eventually, it would be useful to perform a field evaluation of the QMR-DT system, similar to the field testing of QMR described in [16].

4.7 Intended Use of QMR

We used QMR in this study in a manner different from that intended by the system's developers. Specifically, QMR is intended to be used by a physician in an interactive mode [17]. Our use of QMR was limited to applying the QMR diagnostic algorithm once to each set of positive and negative findings. We did not provide the algorithm with additional positive or negative findings based on queries that can be generated by the algorithm. The developers of QMR report that, even after all the positive findings for a case have been entered, the addition of negative findings (to the set of negative findings entered initially) during an interaction with a clinician can increase QMR's diagnostic accuracy [4].

Similar limitations, however, apply to the use and evaluation of the algorithms we implemented, such as S. Within a decision–theoretic framework, it is also possible to analyze a case in an interactive mode. Specifically, given a utility model, we can use value–of–information analyses to guide selection of additional information and refine a disease hypothesis. We are developing approximate methods to compute non-myopically the value of new information given an evidence set [18]. Techniques such as these may allow hypotheses to be iteratively generated and refined in a manner similar to the intended use of QMR [4].

4.8 Shortcomings of the QMR-DT Belief Network Model

Because the current QMR-DT KB is a straightforward reformulation of the INTERNIST-1 KB, both KBs suffer from many of the same shortcomings. The developers of INTERNIST-1 cite several of these deficiencies in [6]: a lack of temporal modeling, a lack of representation of degree of severity, a lack of anatomic knowledge, and an absence of a representation of intermediate pathophysiologic states. Moreover, some of the assumptions we made initially in our probabilistic model, as we discussed in Part I, Section 2.1, may be responsible for the diagnostic inaccuracies of S on the SAM cases. It will be important to investigate which of the assumptions have the most effect on the performance of the system. We believe that the performance will improve when we add dependencies between findings and restructure the KB causally to more correctly model findings that predispose to disease. Also, adding dependencies between diseases probably would improve the performance of the QMR-DT model on multiple-disease cases.

5. Conclusion

The results that we report in this study suggest that the rank-ordering performance of our current probabilistic reformulation of QMR is comparable to that of QMR on cases of the level of difficulty of SAM continuing medical education materials. In addition, it appears that the QMR-DT model is not overtly sensitive to uniform prior probabilities of disease. The model is, however, sensitive to the values of the uniform leak probabilities on findings (the probability that a finding occurs in the absence of any disease in the QMR KB) that we used in this study. We presented evidence showing that the S algorithm produces estimates of posterior marginal probabilities of diseases that are close to the posterior marginal probabilities of diseases implied by the QMR-DT model on the SAM cases. Our sensitivity analysis of the algorithm indicates that on the SAM cases, the S algorithm relies more heavily on the self-importance updating heuristic than it does on the heuristic-importance set.

Since QMR-DT uses a formal probabilistic representation of knowledge, we are able to make explicit each of the assumptions in the model. We plan to test the QMR-DT system further, to investigate those assumptions in the model that are most crucial to system performance. Because QMR-DT is a probabilistic system, we can eventually combine the output of the system with a utility model to create recommendations for cost-effective test ordering and decision-theoretic therapy planning.

The principle result of this study is that we were able to reformulate a large heuristic KB into a probabilistic system that achieved diagnostic accuracy comparable to that of QMR in a laboratory evaluation. Over 20 person years were devoted to building the QMR KB; thus, we saved a substantial amount of time by building QMR-DT as a reformulation of QMR. Having explicitly noted each of the assumptions in the current QMR-DT model, we are now able to begin to evaluate their consequences.

Acknowledgments

We are grateful to Randolph Miller for providing us with the INTERNIST-1 KB and for collaborating with us on the QMR-DT project. We thank Edward Rubinstein from Scientific American Medicine for providing us with test cases for our research. Ross Shachter and Mark Peot provided insight on the likelihood-weighting algorithm. Daniel Bloch assisted with the statistical analyses in this paper. Lyn Dupré provided valuable comments on drafts of this paper. We thank the reviewers for their suggestions to improve this report.

This work was supported by the National Science Foundation under Grant IRI-8703710, the U. S. Army Research Office under Grant P-25514-EL, the National Library of Medicine under Grant R01LM04529, and the National Center for Health Services Research Grant T2HS00028. Computing facilities were provided by the SUMEX-AIM Resource under National Library of Medicine Grant LM05208.

Appendix: Notation and Abbreviations

1. Algorithms

TB	Tabular Bayes' algorithm. An algorithm that uses Bayes' rule under the assumptions of (1) mutually exclusive disease hypotheses and (2) conditional independence of findings given any disease.
ITB	Iterative tabular Bayes' algorithm. A heuristic algorithm that applies TB successively to various subsets of the set of findings.
S	A likelihood-weighting simulation algorithm that uses two heuristics: an heuristic-importance set from ITB and self-importance sampling.
S2	A second run of the S algorithm.
S/NSI	An algorithm identical to S, except that it does not use self-importance sampling.
S/NITB	An algorithm identical to S, except that it does not use an heuristic-importance set generated by ITB.

2. Knowledge base (KB)

d_i	A disease in the KB
f_j	A finding in the KB
F	A set of findings that are observed
F^+	A set of positive findings that are observed
F^-	A set of negative findings that are observed
$\|F^+\|$	The number of elements of F^+
$\|F^-\|$	The number of elements of F^-
H	A hypothesis of diseases, in which each disease is assigned a value of *present* or *absent*

3. Correlation

$P_A(d^+{}_{B(i)} \mid F)$	The marginal posterior probability that algorithm A assigns to the ith-ranked disease of the posterior distribution from algorithm B.

References

1. Cooper GF. The computational complexity of probabilistic inference using Bayesian belief networks. Artificial Intelligence 1990; 42: 393–405.

2. Rubenstein E. Personal communication, 1990.

3. Miller RA. Personal communication, 1989.

4. Miller RA. Personal communication, 1990.

5. Miller RA, Masarie FE Jr. The demise of the "Greek Oracle" model for medical diagnostic systems. Methods of Information in Medicine 1990; 29: 1-2.

6. Miller RA, Pople HEJ, Myers JD. Internist-1: An experimental computer-based diagnostic consultant for general internal medicine. New England Journal of Medicine 1982; 307: 468-76.

7. Miller R, Masarie FE, Myers JD. Quick medical reference (QMR) for diagnostic assistance. MD Computing 1986; 3: 34-48.

8. Ott L. *An Introduction to Statistical Methods and Data Analysis,* Boston, MA: PWS-Kent Publishing Company, 1988.

9. Wyatt J, Spiegelhalter DJ. Evaluating medical expert systems. Medical Informatics 1990; 15: 207-217.

10. Bankowitz RA. User variability in abstracting and entering printed case histories with QUICK MEDICAL REFERENCE (QMR). In: Stead. WW, ed. *Proceedings of the Eleventh Annual Symposium on Computer Applications in Medical Care.* Los Alamitos, CA: IEEE Computer Society Press, 1987: 68-73.

11. Henrion M. Towards efficient probabilistic diagnosis in multiply connected networks. In: Oliver RM, Smith JQ, eds. *Influence Diagrams, Belief Nets and Decision Analysis.* Chichester: Wiley, 1990: 385-407.

12. Cooper GF. *Bayesian Belief-Network Inference Using Recursive Decomposition. Knowledge Systems Laboratory Memo no. KSL-90-05.* Stanford CA: Stanford University, 1990.

13. Pearl J. Evidential reasoning using stochastic simulation of causal models. Artificial Intelligence 1987; 32: 245-57.

14. Shachter RD, Peot M. Simulation approaches to general probabilistic inference on belief networks. In: Henrion M, Shachter R, Kanal LN, Lemmer JF, eds. *Machine Intelligence and Pattern Recognition: Uncertainty in Artificial Intelligence 5.* Amsterdam: North-Holland Publishing Company, 1990: 221-31.

15. Shwe MA, Cooper GF. An empirical analysis of likelihood-weighting simulation on a large, multiply connected medical belief network. Computers and Biomedical Research 1991; to appear.

16. Bankowitz RA, McNeil MA, Challinor SM, Parker RC, Kapoor WN, Miller RA. A computer-assisted medical diagnostic consultation service: Implementation and prospective evaluation of a prototype. Annals of Internal Medicine 1989; 110: 824–32.

17. Miller RA, McNeil MA, Challinor SM, Masarie FEJ, Myers JD. The INTERNIST-1/QUICK MEDICAL REFERENCE project—Status report. Western Journal of Medicine 1986; 145: 816-22.

18. Heckerman DE, Horvitz EJ, Middleton B. *An Approximate Nonmyopic computation for Value of Information. Knowledge Systems Laboratory Memo no. KSL-91-15.* Stanford CA: Stanford University, 1991.

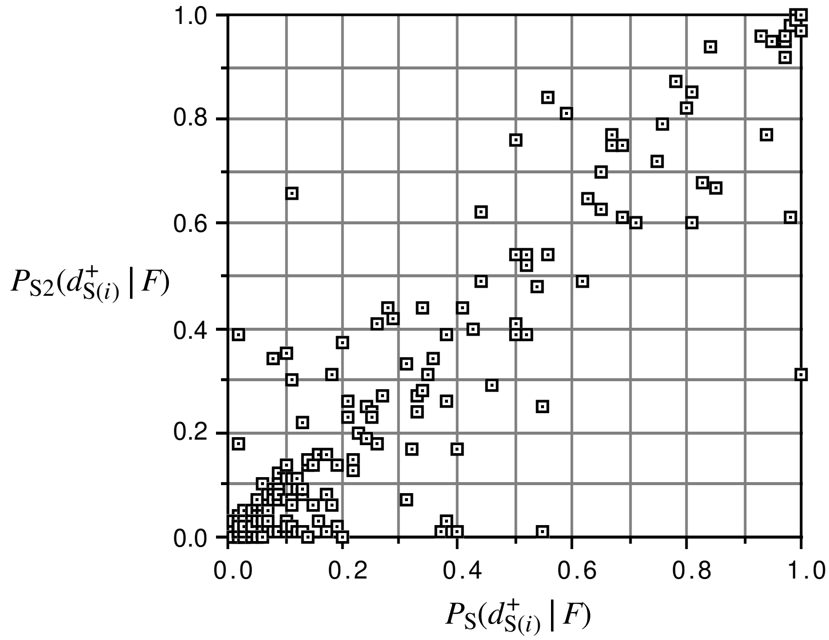

Fig. 1

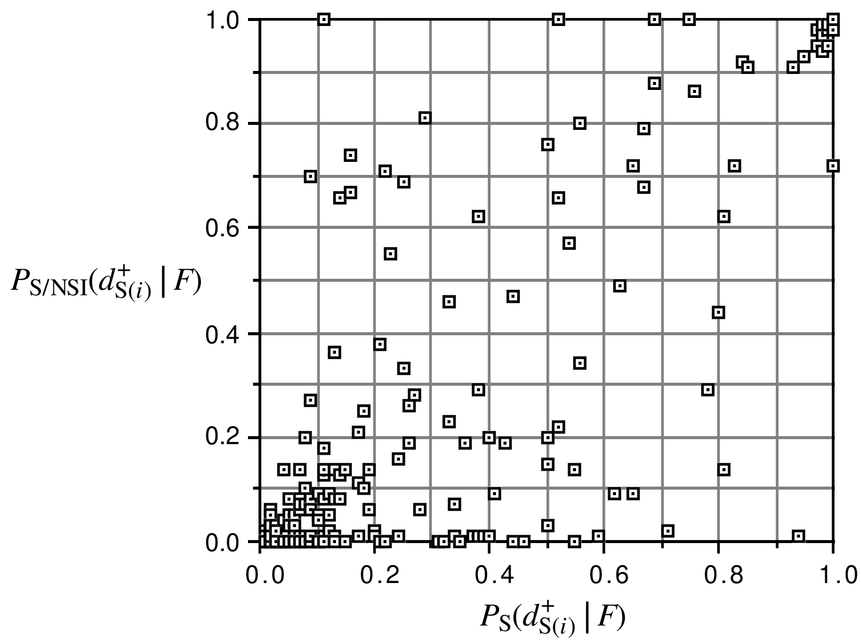

Fig. 2

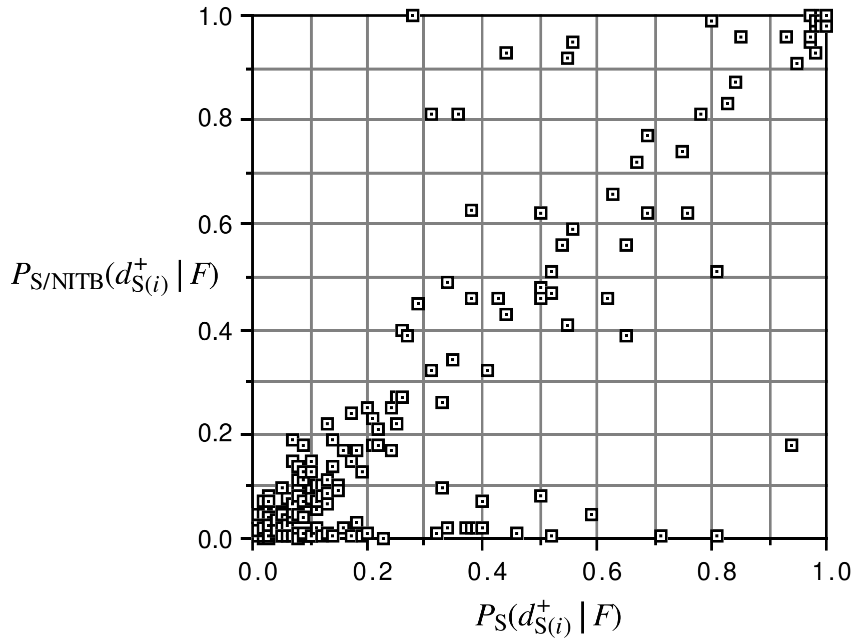

Fig. 3

Table 1 Diagnoses and number of findings for the single-disease Scientific American Medicine (SAM) cases (N = 23) used in the evaluation

SAM case number	Diagnosis	$\lvert F^{+}\rvert$[a]	$\lvert F^{-}\rvert$[b]
1	acute myocardial infarction	18	11
6	ulcerative colitis	11	15
15	chronic active hepatitis	16	9
20	systemic lupus erythematosus	15	22
22	acute myocardial infarction	24	8
23	celiac sprue	18	18
25	acute pyelonephritis	26	30
27	pulmonary tuberculosis	12	8
28	mitral-valve prolapse	20	8
29	Legionnaires' pneumonia	22	8
30	idiopathic thrombocytopenic purpura	17	21
31	primary transient sarcoidosis	24	19
33	nephrolithiasis	25	30
34	diffuse esophageal spasm	17	45
35	Alzheimer's disease	16	18
37	idiopathic pericarditis	31	29
40	giant cell arteritis	12	29
42	acute myocardial infarction	17	25
46	ulcerative colitis	18	21
47	acromegaly	16	28
50	polycythemia vera	12	19
51	thyroid papillary carcinoma	10	20
53	aortic dissection	17	26
	mean	18	20

[a] $\lvert F^{+}\rvert$ is the number of positive findings in the case
[b] $\lvert F^{-}\rvert$ is the number of negative findings in the case

Table 2 Ranks assigned to the reference diagnoses of the 23 SAM cases

	Algorithm					
SAM case number	QMR	TB	ITB	S	S/UD	S/UL
1	6	1	1	1	1	1
6	2	2	1	2	2	2
15	1	1	1	2	2	1
20	1	1	1	1	1	1
22	1	1	1	1	2	1
23†	—(1)	5 (1)	20 (1)	103 (1)	4 (1)	216 (1)
25	3	1	2	1	2	6
27	1	1	3	1	1	1
28	1	2	1	1	1	1
29	3	4	11	9	6	106
30	5	2	3	7	17	36
31	12	9	11	24	166	255
33	2	2	17	2	1	1
34	1	6	12	4	4	445
35	1	1	3	1	2	2
37	2	17	2	2	7	8
40	1	1	1	1	1	352
42	4	1	3	2	2	1
46	1	1	1	1	1	1
47	1	1	1	1	1	1
50	1	1	2	1	1	1
51	2	2	5	57	22	30
53	3	1	1	1	1	1

Key:

— Reference diagnosis not ranked

† In case 23, we identified retrospectively an intermediate pathophysiologic state of malabsorption. The rank of malabsorption appears in parentheses for each algorithm.

Table 3 Summary of the ranks assigned to the reference diagnoses of SAM cases (N = 23)

	Algorithm											
Summary statistic	QMR	(%)	TB	(%)	ITB	(%)	S	(%)	S/UD	(%)	S/UL	(%)
Number in top 1	11	(48)	13	(57)	10	(43)	12	(52)	10	(43)	12	(52)
Number in top 5	21	(91)	20	(87)	18	(78)	18	(78)	18	(78)	14	(61)
Number in top 10	21	(91)	22	(96)	18	(78)	20	(87)	20	(87)	16	(70)
Number in top 20	22	(96)	23	(100)	23	(100)	20	(87)	21	(91)	16	(70)

Table 4 Values of the test statistic, T, for a two-sided Wilcoxon signed-rank test comparing the rank ordering generated by algorithm S with the rank orderings generated by five other algorithms

	Algorithm				
	QMR	TB	ITB	S/UD	S/UL
Wilcoxon T:	15	13.5	39	59	**14.5**[b]
$T(0.05,n)$[a]	13	10	21	25	17
n	11	10	14	15	13

a $T(0.05,n)$ is the critical value for the T statistic at the $p = 0.05$ level of significance for n pairs of observations with nonzero difference.

b When the computed value of T is less than $T(0.05,n)$, we can reject the null hypothesis (at the $p = 0.05$ level) that the rank ordering of S is identical to that of another algorithm.

Table 5 Correlation coefficients comparing the estimates of the posterior marginal probabilities of diseases generated by three algorithms to the posterior distribution of S on the SAM cases

	Correlation coefficient		
SAM case number	*r*(S, S2)	*r*(S, S/NSI)	*r*(S, S/NITB)
1	0.96	0.83	0.99
6	1.00	0.97	0.95
15	0.92	0.78	0.67
20	1.00	1.00	1.00
22	0.61	0.30	0.63
23	0.87	0.89	0.99
25	0.97	0.87	0.68
27	0.94	0.96	1.00
28	0.97	0.94	0.99
29	0.91	0.97	0.96
30	0.90	1.00	1.00
31	0.94	0.37	0.79
33	0.98	0.79	0.92
34	0.98	0.51	0.97
35	1.00	1.00	0.99
37	0.82	0.65	0.67
40	1.00	0.97	1.00
42	0.99	0.98	0.99
46	1.00	1.00	1.00
47	1.00	0.76	0.90
50	0.99	0.97	0.96
51	1.00	1.00	1.00
53	1.00	0.91	0.99
pooled[a]	0.93	0.81	0.89

a The pooled correlation coefficient is computed from the pairs $(P_S(d^+{}_{S(i)} \mid F), P_B(d^+{}_{S(i)} \mid F))$ for $1 \le i \le 10$ for all of the SAM cases.

Fig. 1 A plot of the posterior estimates of S2 as a function of the corresponding top 10 estimates from S, pooled from the 23 SAM cases. ($r = 0.93$.)

Fig. 2 A plot of the posterior estimates of S/NSI as a function of the corresponding top 10 estimates from S, pooled from the 23 SAM cases. ($r = 0.81$.)

Fig. 3 A plot of the posterior estimates of S/NITB as a function of the corresponding top 10 estimates from S, pooled from the 23 SAM cases. ($r = 0.89$.)