Research paper ■

A Temporal Analysis of QMR

CONSTANTIN F. ALIFERIS, MD, MS, GREGORY F. COOPER, MD, PHD,
RANDOLPH A. MILLER, MD, BRUCE G. BUCHANAN, PHD,
RICHARD BANKOWITZ, MD, NUNZIA GIUSE, MD, MS

Abstract **Objective:** To understand better the trade-offs of not incorporating explicit time in Quick Medical Reference (QMR), a diagnostic system in the domain of general internal medicine, along the dimensions of expressive power and diagnostic accuracy.

Design: The study was conducted in two phases. Phase I was a descriptive analysis of the temporal abstractions incorporated in QMR's terms. Phase II was a pseudo-prospective controlled experiment, measuring the effect of history and physical examination temporal content on the diagnostic accuracy of QMR.

Measurements: For each QMR finding that would fit our operational definition of *temporal finding*, several parameters describing the temporal nature of the finding were assessed, the most important ones being: temporal primitives, time units, temporal uncertainty, processes, and patterns. The history, physical examination, and initial laboratory results of 105 consecutive patients admitted to the Pittsburgh University Presbyterian Hospital were analyzed for temporal content and factors that could potentially influence diagnostic accuracy (these included: rareness of primary diagnosis, case length, uncertainty, spatial/causal information, and multiple diseases).

Results: 776 findings were identified as temporal. The authors developed an ontology describing the terms utilized by QMR developers to express temporal knowledge. The authors classified the temporal abstractions found in QMR in 116 temporal types, 11 temporal templates, and a temporal hierarchy. The odds of QMR's making a correct diagnosis in high temporal complexity cases is 0.7 the odds when the temporal complexity is lower, but this result is not statistically significant (95% confidence interval = 0.27–1.83).

Conclusions: QMR contains extensive implicit time modeling. These results support the conclusion that the abstracted encoding of time in the medical knowledge of QMR does not induce a diagnostic performance penalty.

■ **JAMIA.** 1996;3:79–91.

A frequently encountered heuristic in the development of medical decision support systems (MDSSs) is the implicit or abstracted modeling of time.[1,2] This heuristic has significant implications for the design, implementation, and application phases of any given system that employs it. For the purposes of this paper, we define *explicit time* to be the handling of time that incorporates the three following components: 1) a time model, with well-defined fundamental temporal entities and properties, examples of which include temporal primitives (points, intervals, processes) and a specific structure for time (a set of temporal properties such as direction, finiteness, and continuity); 2) a language for expressing the association of entities (objects, relations) with the time model (for example, events occurring within intervals, facts

Affiliations of the authors: Section of Medical Informatics and the Intelligent Systems Program (CFA, GFC), and the Department of Computer Science (BGB), University of Pittsburgh, Pittsburgh, PA; the Informatics Center (RAM, NG), Vanderbilt University, Nashville, TN; and the University Hospital Consortium (RB), Oak Brook, IL. At the time of the study, Drs. Miller and Giuse were at the University of Pittsburgh, Pittsburgh, PA.

Presented in part at the 18th Annual Symposium on Computer Applications in Medical Care, November 1994, Washington, DC, and at the 8th World Congress on Medical Informatics, July 1995, Vancouver, BC, Canada.

Correspondence and reprints: C. F. Aliferis, MD, MS, Section of Medical Informatics, B50A Lothrop Hall, 190 Lothrop Street, University of Pittsburgh, Pittsburgh, PA 15261. e-mail: cons@alpha.smi.med.pitt.edu

Received for publication: 7/20/95; accepted for publication: 9/12/95.

being true or false for part or all of a time period); and 3) a set of appropriate inference rules that exploit knowledge about time and temporal associations to solve interesting problems (for example, if an event occurs before an interval, then it does not co-occur with any event following that interval).[3-5]

This type of time modeling is in contrast with what we call *implicit time*, which is characterized by:

1. Building temporal representations and inferences into propositional statements.
2. Utilizing ordinary atemporal inference procedures to reason with the propositional statements.

Typically, but not always, the user of the system is the abstractor of information (that is, the one who will provide the system with the truth or falsity of a proposition by utilizing his or her own temporal reasoning capabilities).

The research reported here intends to investigate the following hypothesis: *For specific MDSSs and/or domains, abstracted time can achieve adequate performance, and thus by using it, one can avoid explicit time modeling and its associated costs.* There are a number of additional interesting questions associated with this conjecture:

1. What constitutes an appropriate collection of abstracted (atemporal) knowledge representations, corresponding to the domain to be modeled?
2. Are there specific temporal entities that are crucial to MDSSs' performance? What is the proper level of description of those entities?
3. How would these results be useful for systems that operate in automated mode (i.e., when a human operator mapping patient information to system temporal abstractions is not available)?

We focus our attention on the problem of diagnosis, in the domain of general internal medicine (an important and definitely non-trivial medical problem) and a particular system designed to solve it, Quick Medical Reference (QMR). QMR employs implicit time modeling and has been successfully evaluated against humans and similar systems, and carefully maintained since the inception of its precursor INTERNIST-I.[6-9]

In particular, we wanted to examine the importance and effect of representing time implicitly on the performance of the system. In a classic statistics framework, this amounts to building a research design around the null hypothesis that the lack of explicit time in QMR does not cause decreased diagnostic performance, and trying to reject or accept the hypothesis. Part II of our study examines this hypothesis, as described below.

An equally important task is the explanation of *why* the hypothesis is refuted or corroborated by experimental data; in other words, we must identify those characteristics of QMR's implicit time handling that are responsible for its temporal robustness or insufficiency. To provide such an analysis we need to first understand better the nature of implicit time in QMR. Part I of the study seeks to provide such an understanding by studying the temporal abstractions found in QMR.

Background

In a review of temporal modeling in MDSSs, Kahn[5] proposes an empirical classification that places systems in a spectrum of categories having at one extreme *temporal ignorance* (equivalent to out definition of implicit time), and at the other extreme *integrated systems* (i.e., systems based on a multitude of temporal models working in coordination to solve a particular task). More specifically, he demonstrates how the earlier systems avoided the need for explicit knowledge representation and reasoning (KRR) by incorporating temporal information into ordinary atemporal formalisms. For instance, the INTERNIST-I system[6] would ask questions of the type "Did the patient have a history of disease x?," which clearly corresponds to an abstracted handling of time. It is obvious that the system's developers were operating on the assumption that the user of the system would abstract relative data from historical observations and provide it to the program. The same approach was followed in a number of influential systems such as MYCIN, PIP, DXplain, CASNET, and ABEL.[10]

The popularity of abstracted time in MDSSs can be attributed to factors that include the following:

1. Because temporal information is reduced to fewer or more abstracted variables, complex evidence is grouped together, so that the resulting model is economical (or restricted, depending on one's viewpoint) as far as complexity of evidential support is concerned.
2. Temporally implicit models require fewer parameters than do their temporally explicit counterparts. Thus, explicit models generally require more knowledge and data acquisition. Equivalently, if we make assumptions about unknown model pa-

rameters, then temporally implicit models require less assumptions than do explicit ones.

3. Temporally implicit models rely on some external mechanism (typically a human operator) to provide the MDSS with the truth or falsity of an abstracted temporal proposition, or alternatively, require an automated temporal abstraction mechanism (usually coupled with a historical or fully temporal database). Thus, they exploit human reasoning capabilities or decompose the temporal reasoning task into simpler subtasks.

4. In addressing artificial intelligence (AI) problems, more expressive power generally means less computational tractability (and the opposite). Thus, we would expect that the ability to express temporal aspects of the problem domain is naturally followed by increased computational complexity.[11]

5. Abstracted time was easier to model, especially in the earlier years of medical AI, when there was a relative lack of well-defined temporal models. Recent advances in temporal modeling have reduced the importance of this factor for modern MDSSs. Such advances include temporal logics,[3,4] connectionism methods, Markov decision processes, temporal belief networks/influence diagrams,[12,13] specific MDSSs that deal with dynamic domains,[12,14–18] temporal databases, temporal database query languages,[19] and automated temporal abstraction methods.[20]

These arguments suggest that there are important trade-offs between explicit and abstracted time modeling. In other words, we need to examine *why* and *how* important is thc ability to reason explicitly (as contrasted to an abstracted manner) about temporal processes and entities. The necessity of explicit time has been only partially explored in the medical AI and medical informatics literature, especially with respect to *quantifying* this importance. To our knowledge, there has not yet been a formal theoretical or empirical analysis of the trade-offs between explicit and implicit time for any realistic medical domain or MDSS.

The basic arguments that have been offered in favor of the importance of explicit time in MDSSs are:

1. The *epistemological argument*: observations of physicians' diagnostic and therapeutic problem solving suggest that temporal models of normal and abnormal processes are used, and intricate temporal abstractions are created and employed to generate and validate (or rule out) competing hypotheses. Additionally, physicians are able to utilize temporal planning for either diagnosis (e.g., "watchful expectancy") or therapy.[21]

2. The *linguistic argument*: analyses of discharge summaries and other medical texts indicate an impressive amount of temporal reasoning.[14]

3. *Pragmatics arguments:*

- The *temporal domain argument*: certain medical domains are based on the premise of a time-evolving process, and explicit time is fundamental for them [characteristic examples include the protocol-based therapy management, intensive care unit (ICU) real-time monitoring and intervention, and signal processing as in electrocardiographic (EKG) and electroencephalographic (EEG) interpretation].[15,22]
- The *failure analysis argument*: evaluation of MDSS diagnostic performance shows that some failures to reach the proper diagnosis can be attributed to a lack of temporal capabilities.[6,16]

The epistemological and linguistic arguments are purely *descriptive* and do not in themselves prove the importance of explicit temporal reasoning in MDSSs. The temporal domain argument is true, but refers to clearly defined narrow types of MDSSs. There remains an open question about the importance of explicit time in many areas of medicine, such as the domains of INTERNIST-I and MYCIN. These systems' need for explicit time could be substantiated by the failure analysis argument, in the sense that, ceteris paribus, if explicit time accounts for a substantial number of diagnostic failures and the problems cannot be fixed using implicit time models, then we can conclude that explicit time is indeed necessary.

Through the present time, support for the failure analysis argument comes in the form of anecdotal evidence, rather than from experiments designed to investigate the validity of this hypothesis. One often-cited example is the 1982 *New England Journal of Medicine* (*NEJM*) evaluation of INTERNIST-I, which, on the basis of three cases (of a total of 19 diagnostic problems), indicates that failure to represent explicit time caused diagnostic errors. But the 99% confidence limits of 3/19 (16%) are between 2% and 47%, suggesting that no strong conclusion should be reached from these data regarding the effects of not representing explicit time. Even more importantly, the cases were not representative of the average encountered clinical case, since they were *NEJM* clinicopathological conference cases that were selected on the basis of being very challenging.[6]

The previous discussion indicates a need for further investigation and quantitative analysis of the importance of explicit time in MDSSs. For certain MDSS domains, this need seems well justified by the nature of the domain (i.e., the nature of the entities represented is so deeply temporal, that either we could not reason about it without taking into account time, or it would be grossly ineffective to utilize some implicit/abstracted form of KRR). These domains/tasks include:

- protocol therapy management,[5,20]
- biomedical signal processing,[22]
- "deep" causal models of diseases/physiology, which are modeled as dynamic systems,[23]
- ICU decision support,[15] and
- human growth assessment.[17]

In general terms, the characteristics of domains that seem to *require* explicit time are:

- optimization according to some time-dependent utility function,
- very small temporal scale,
- highly time-critical interventions, and
- the need for high precision in the identification of temporal patterns.

Methods

Temporal Analysis of QMR's Terms

We devised a series of variables that correspond to what previous theoretical and empirical work suggests are important temporal reasoning and representation attributes.[3-5,11] These were used by the first author to classify each finding in QMR as *temporal* or atemporal, based on the following criterion. We classify a QMR finding as temporal if *any* of the following is true: 1) it contains an *explicit* reference to either time points/intervals or units, 2) it refers to temporal relationships/reasoning, 3) it describes events or facts in some temporal context, 4) it refers to processes occurring over time (explicitly stable/evolving, or in sequence/overlapping), or 5) it mentions specific patterns (temporal or spatiotemporal).

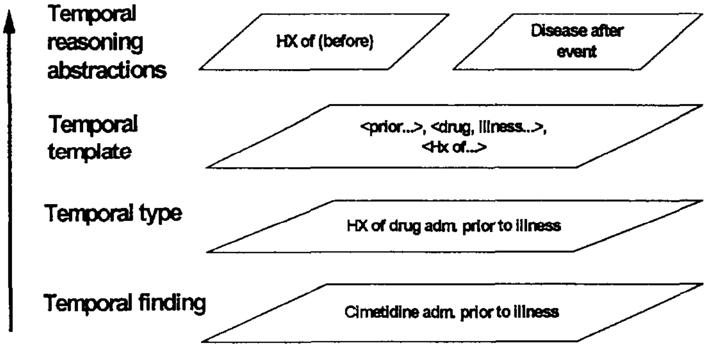
Temporal reasoning abstractions
HX of (before)
Disease after event
Temporal template
<prior...>, <drug, illness...>, <Hx of...>
Temporal type
HX of drug adm. prior to illness
Temporal finding
Cimetidine adm. prior to illness

Figure 1 Levels in the abstraction process. Hx = history.

We additionally noted, from their representation in the QMR knowledge base (KB), the *QMR finding type* (history, physical, simple–inexpensive laboratory, intermediate laboratory, advanced–expensive laboratory) and the QMR finding *importance* (i.e., the "import" value of QMR, which indicates the "need for a finding to be explained diagnostically if found"[6]), for all findings, regardless of temporal nature. We developed abstractions over the temporal findings (temporal types, temporal templates) and developed a temporal ontology for QMR, and examined temporal reasoning in QMR, in an incremental fashion, refining the abstractions as new QMR findings were examined. Temporal types correspond to simple abstractions over QMR findings. A *template* is an abstraction over temporal types designed to capture in a concise manner the temporal aspects of types. A temporal type typically is an instantiation of only one template, although a few types can be viewed as specializations of more than one template. Templates can be used to generate types, although some instantiations of the templates may not belong in the original types found in QMR. A template has the following structure:

[{temporal relation}, {entities}, {qualifications}]

The *temporal relation* is a set of temporal knowledge or reasoning statements characterizing the template. *Entities* is a set of non-temporal or temporal primitives that serve as arguments to the temporal relation. Finally, the *qualifications* set provides specific details as to what the particular nature of the various types captured by the template might be. The following example (partial description of template 10) illustrates these concepts:

[{worsening, improvement, rapidly progressing}, {disease, abnormal finding, symptom}, {Hx, recent}]

From this example template we can derive various temporal types, for instance: "abnormal finding rapidly progressing" (type 75), or "Hx (history) of recent worsening/improvement of abnormal finding/disease" (type 25). We can also derive types *not* found in QMR, but which are reasonable generalizations/variations of the existing types (e.g., "Hx of worsening symptom"). This occurs because in QMR the user cannot enter *all* findings with a temporal qualification, even if such a qualification exists for some findings. Whereas the temporal types provide the specific temporal abstractions employed by QMR, the templates serve as a *summary* and a *generalization* of those abstractions. A third concept, which we call *reasoning types,* denotes fundamental relations and other properties that can be combined to form temporal types (for examples of temporal types and reasoning types, see Results sections 1.2–1.4).

To ensure consistency in the categorization of temporal QMR findings (according to temporal type and reasoning type), the following procedure was followed: First, temporal findings were identified and separated from the rest of the QMR findings. Temporal types were developed from the temporal QMR findings. Second, values for the variables (i.e., temporal characteristics) for *each temporal type* were assigned. Due to the limited number of types (<120), consistency checks (with previously established temporal types) were easier and less error-prone to carry out than they would be with the full set of temporal QMR findings (776 in total). Third, after the types had been characterized, individual findings were categorized as belonging to any specific temporal type. As a consequence, each finding would inherit the temporal attribute assignments of the corresponding abstract type. Fourth, each finding was examined for differences with the type it belonged to (due to the abstraction process), and the necessary adjustments were made to the deviating attributes of the individual findings. Finally, templates and temporal reasoning types were abstracted and classified empirically. Figure 1 illustrates the abstraction process from actual findings to temporal types, templates, and reasoning types.

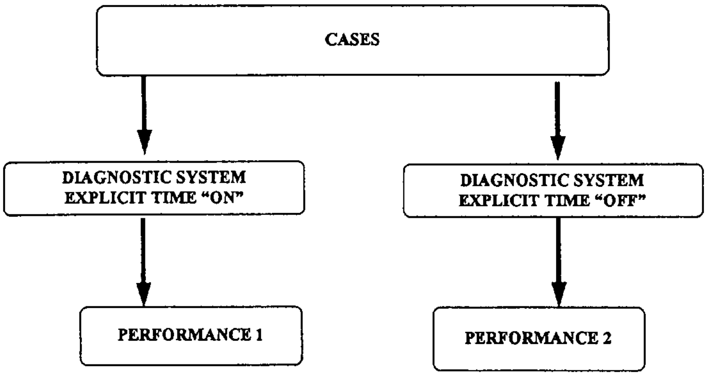
CASES
DIAGNOSTIC SYSTEM EXPLICIT TIME "ON"
DIAGNOSTIC SYSTEM EXPLICIT TIME "OFF"
PERFORMANCE 1
PERFORMANCE 2

Figure 2 An idealized experiment.

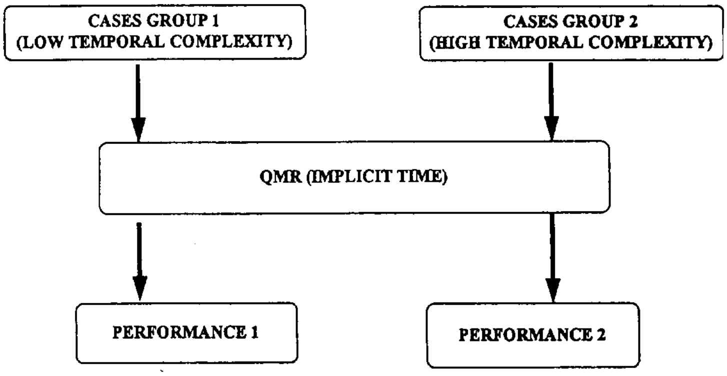
CASES GROUP 1 (LOW TEMPORAL COMPLEXITY)
CASES GROUP 2 (HIGH TEMPORAL COMPLEXITY)
QMR (IMPLICIT TIME)
PERFORMANCE 1
PERFORMANCE 2

Figure 3 A modified experiment. QMR = Quick Medical Reference.

Standard descriptive statistics were computed for all variables. Bivariate associations of temporal attributes with the temporal classification/import/QMR-type of findings were examined with likelihood ratio (G^2) tests of independence, Kendall's tau, and the gamma coefficient (for ordinal variables). The association of import with temporal classification was further examined with the previous statistics controlling for possible confounders.[24]

Effects of Lack of Explicit Time on Diagnostic Performance

Ideally, we would like to test the following (null) statistical hypothesis: *Lack of explicit time in QMR does not cause decreased diagnostic performance when compared with the case where explicit time is employed.* Figure 2 shows an idealized experiment built around a posttest design[25] in which the same group of cases is presented to the system. Assuming that the diagnostic system has explicit time that can be turned on and off at will, diagnoses are performed twice, once with explicit time being active, and once with explicit time being inactive. The performance in the first case is compared with that in the second one. Unfortunately, this ideal experiment is unattainable. There is no MDSS employing explicit time that operates with a scope comparable to that of general internal medicine. Nor is temporal reasoning typically implemented in a manner that can be turned on and off. Since modifying QMR to incorporate an explicit time model is equally infeasible for the purposes of this study,[26] we designed a modified version of the previous experiment, represented in Figure 3.

Table 1 ■

The Most Frequent Temporal Types

1. Hx* of syndrome/disease [11.2%]
2. Hx of drug administration prior to current illness [8%]
3. Improvement/worsening of function after/during test/medical procedure/state [7.3%]
4. Hx of familial disease/behavior [7.2%]
5. Abnormal/normal finding/syndrome after drug/medical procedure [4.4%]
6. Hx of exposure to animals/factors [3.6%]
7. Hx of recent medical procedure [3%]
8. Hx of recent exposure to factor environment/food/behavior [2.7%]
9. Increased/decreased rhythm/rate/speed [2.4%]
10. Measurement per unit of time > C† [2.3%]

*Hx = history.
†C = constant value.

This is a pseudo-prospective design, where we first defined a measure of temporal content for patient records. Then we collected a set of patient cases and separated them in two groups: one of high and one of low temporal complexity. We presented each group of cases to the diagnostic system and derived a differential diagnosis for each case. Utilizing an appropriate diagnostic performance definition, we derived a performance measure for each group. If the two groups differed only in their temporal contents, then we concluded that any difference found in the performance measures would be attributed to the inability of the system's implicit time mechanism to cope with the temporal information found in the cases.

We used 105 cases from the most recent formal evaluation of QMR[27] (each consisting of history, physical, initial laboratory tests, and discharge summary and diagnoses). One of us (RB) was the primary investigator in that study. The coding of the patient information was done by experienced QMR users under the supervision of the last two authors (RB, NG). The patients were considered to be representative of the patients admitted to a large university hospital, since they were consecutive, unselected, patients presenting to a university hospital covering a large urban area.

The key concept in our experimental design was to make certain that the two patient groups indeed differed only in temporal content, and were similar in terms of other properties that were suspected or known to be sources of diagnostic difficulty or even failure. First, we note that the history and physical examination text of each patient record was separated into a number of individual pieces of information (POIs). A POI was defined as the smallest piece of clinically relevant information that could be meaningful if stated in the given document context. Thus, a POI could be either a stand-alone statement or a qualification of a previously established statement. To establish comparability between the two groups, we measured a set of potentially confounding variables (in a blinded fashion with respect to outcome), which were: rareness of the primary diagnosis; case length; presence of uncertainty (as percentage of uncertain POIs); use of spatial and causal information; number of diseases in the gold standard (GS)—that is, discharge-verified diagnoses for the patient case; and levels of reasoning involved. Note that it is important to maintain a prospective design, to avoid a case–control setup (i.e., trying to identify the temporal differences between the cases for which the system had a high performance vs the cases for which it had a low diagnostic performance), and the associated potential biases with respect to identifying the risk factor, and establishing case–control comparability.[28] Other important considerations in the execution of this design are:

1. Temporal content assessment: Each POI in the history and physical (H&P) text was characterized as temporal or not, based on the same criteria used to classify QMR findings (see Methods, section 1). For each POI, the values of the confounder attributes were assessed. The percentage of temporal POIs, divided by the total number of POIs in the case, constituted our measure of temporal content for that case. For each POI, the temporal attributes utilized in the assessment of explicit time were evaluated and summarized for each case. We utilized principal components analysis to identify summary linear combinations of those attributes as more detailed metrics of the case temporal content. Similar measures of complexity and temporal content were assessed for the QMR encodings for each case. Finally, we identified temporal types in the raw clinical case descriptions that exceeded the expressive capacity of the QMR abstractions. The assignment of a value to each temporal attribute (for both the patient H&P and the QMR inputs) was done by the first author, who was blinded to the case outcome. The confounding variables were defined as follows (most based on the identification of POIs) for each case: the rareness of the primary diagnosis was measured as the prevalence of disease as recorded in the QMR KB (via a quasi-logarithmic prevalence index), the case length as the number of POIs, the presence of uncertainty as the percentage of uncertain POIs, the use of spatial and causal information as the percentage of spatial and causal POIs, respectively, the number of diseases in the GS diagnosis

as such, and, finally, the levels of reasoning involved as the number of distinct levels referred to in the case.

2. Performance assessment: Our criterion was the percentage of cases for which QMR found the primary diagnosis (i.e., the top diagnosis in the discharge summary diagnoses list).

The following matching criterion was used:

- A diagnostic match occurred if and only if the GS primary diagnosis is clinically *close* to one of the q first diagnoses in the QMR differential diagnosis list, where q is a percentage. The primary discharge diagnosis (ICD9 primary diagnosis) was considered to be the GS.
- q is defined to be a percentage of diagnoses from QMR's differential diagnosis list. We experimented with various values for the q parameter (see Methods), and decided to use q = 100% to provide a better balance of sample size between successful and non-successful diagnostic groups (in retrospect, sensitivity analysis shows that our results are insensitive to this parameter for the range of all possible values: 20% to 100%).
- A QMR diagnosis was judged clinically close to the GS if it was either identical to it, synonymous with it, in a significantly overlapping disease category, or at most one level down or up in a recognized clinical classification such as those found in major textbooks of medicine (e.g., Harrison's or Cecil's textbook of medicine).

Cases with no established primary diagnosis were excluded. When the first (primary) diagnosis in the GS differential was asserted in QMR, or given as a finding in QMR, or was not in QMR's KB, the next diagnosis was used as the primary one (with a recursive application of the exclusion and skipping rules). The first author performed all the matches manually. Based on our diagnostic criteria, we had to exclude a number of patient cases, for the following reasons: the diagnosis was a finding in QMR; the diagnosis was not part of the QMR's KB; the cases did not represent a straightforward diagnostic problem (but a therapeutic or "rule out" problem); QMR did not produce a diagnostic list; diagnoses were asserted (i.e., given to the system as fact); or all the necessary information was not available in the patient record. Thus, 35 of 105 cases were excluded from subsequent analyses.

Table 2 ■

Temporal Templates and the Corresponding Numbers of Temporal Types That Each Captures

1. [{migrating}, {finding, symptom}, {Hx,* now}]: 2
2. [{simultaneous}, {findings, symptoms}, { }]: 1
3. [{cardiac-pulse-specific pattern}, { }, { }]: 1
4. [{impending}, {death}, {fear of}]: 1
5. [{single}, {abnormal finding}, { }]: 1
6. [{Hx of}, {finding, symptom, disease, state, syndrome, behavior, causal or evidential events}, {recent, remote, childhood, congenital, or now, by Hx or current information}]: 16
7. [{repetition (implicitly)}, {finding, symptom, syndrome, disease, behavior, events, causal factor, medical procedure, causal or evidential event}, {Hx, chronic, recurrent, paroxysmal, with remission, premature by Dt,† episodic, multiple, seasonal, >*n*/interval, irregular, intermittent}]: 26
8. [{after, with, and, epidemic (i.e., in the context of)}, {abnormal/normal finding, disease, syndrome, abnormal function, symptom, state, improvement/worsening of function/finding/syndrome, //‡ drug, medical procedure, animal, environment, food, factor, behavior, event, exercise, factor presence, factor use}, {current, recurrent, greater than duration, Hx, remote, recent, smaller or equal than duration}]: 32
9. [{during, with}, {abnormal finding, disease, symptom, signs, // period, disease episode, state, activity, decreased measurement}, {Hx, at onset, improving/worsening}]: 11
10. [{change, worsening, improvement, rapidly progressing, maximum severity, worsening followed by improvement, rise and fall, transient, progressive, changing character}, {characteristics, abnormal findings, disease, finding, symptom, lab value, syndrome}, {Hx, recent, at onset, in period of time}]: 19
11. [{change in, age=, acute, paroxysmal, prolonged, increased duration/severity, increase, decrease, continuous, lasting > duration/ <duration, duration of period, measurement per units of time > constant}, {measurement time, finding, symptom, finding, behavior, recovery, healing, rhythm, rate, speed, medical procedure, factor, abnormal/normal function, measurement, drug administration, rate of normal response in diagnostic test}, {Hx, recent}]: 20

*Hx = history.

†Dt = amount of time.

‡A double slash ("//") separates the two parameter lists in relations with two arguments.

Table 3 ■

Temporal Reasoning Type Abstractions

1. Spatiotemporal evolution (i.e., change in time *and* space)
2. Simultaneity (i.e., concurrence of events and/or facts)
3. Pattern recognition (i.e., specific clinically meaningful pattern detection)
4. Temporal projection (i.e., specification of expected events as a result of current actions or states)
5. Event singularity (i.e., specification of non-repetition)
6. History of (i.e., precedence)
7. Current property (i.e., time of reference is "now")
8. Form of progression/onset (i.e., specification of start or evolution of state)
9. Repetition (i.e., periodicity, regularity, recurrence, multiplicity, seasonality, counting, rates, continuity/discontinuity)
10. Temporal location (i.e., succession, coincidence, during-relations)
11. Interval duration
12. Severity change

3. Analysis: All continuous variables were discretized (based on their observed 50th percentile value as a single cutoff point). Odds ratios (ORs) of correct diagnoses were computed between the explanatory variable (i.e., potential confounders, and measures of temporal content) categories.[24] Logit models (using the continuous variable versions and a standard statistical package) and Bayesian models (through the application of the K2 inductive learning algorithm[29]) were built to assess quantitatively the impact of temporal case content on system diagnostic accuracy. The interrelations of temporal content and the rest of the explanatory variables were also examined with respect to diagnostic accuracy. Finally, the confidence profile method (as implemented in the Fastpro software package[30]) was used to derive high-density regions for the univariate ORs of the explanatory variables, assuming uniform prior distributions on the probability of correct diagnosis in the two temporal content categories.

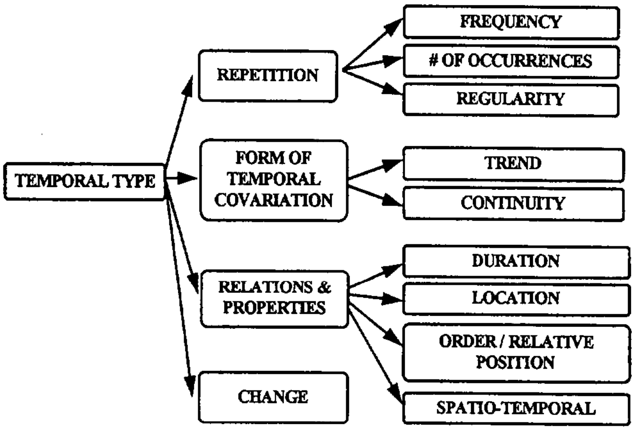

Figure 4 Temporal type abstraction.

Results

Temporal Analysis of QMR's Terms

Ontology

We found that QMR utilizes the following temporal ontology to express temporal findings.

1. Entities:

- Generic: disease, syndrome, finding, symptom, laboratory value, test result, medical procedure, drug, causal factor, diagnostic factor, behavior, function, state, sign.
- Temporal: periods, points of time, seasons, parts of the day, disease intervals, EKG-related intervals, systolic/diastolic periods, units of time, parts of intervals, age.

2. Relationships/properties:

- History of, during, before, after, coincides with, repeating, properties (frequency, speed, rhythm, regularity), duration, specific patterns. Also, Boolean combinations of the above are used to derive more complex propositions.

Temporal types

We constructed a total of 116 temporal types based on QMR findings. Table 1 lists the most frequent ones, together with their frequencies (% of total number of temporal findings). [Another 49 temporal types abstracted over the types presented here are given in reference 31 (but not discussed in this paper, since they are subsumed by the temporal templates). The full list of temporal types can also be found in reference 31.]

Temporal templates

Table 2 contains the descriptions of the 11 templates derived from the temporal types, followed by the number of the temporal types captured by each template. Templates 1 to 5 correspond to a few types, where templates 6 to 11 generalize over many more type instances.

Temporal reasoning abstraction

We identified a total of 12 different temporal reasoning abstractions (as well as combinations of those), and they are shown in Table 3.

Statistical associations

We found several interesting associations among the QMR temporal finding properties. QMR findings that reference temporal units (minutes, hours, etc.) have higher QMR import values than those of findings that do not (G^2 $p < 0.0001$, gamma $= -0.51$ with t-value $= -3.6$). Similarly, when a QMR finding makes an explicit reference to procedures or patterns, the import value is higher than that of findings that do not (G^2 $p = 0.015$ and 0.0005, gamma $= -0.64$ and 0.64 with t-values $= -1.1$ and 3.44, respectively). Overall, however (i.e., in the full set of QMR findings), temporal QMR findings have lower QMR import values than do non-temporal findings (G^2 $p <$ 0.0001, gamma $= 0.65$ with t-value $= 23.2$). Symptoms and signs have less import than do more advanced laboratory findings (G^2 $p < 0.0001$, tau $=$ 0.11 with $p < 0.0001$). At the same time, temporal findings are characterized by smaller values in the QMR type scale of diagnostic sophistication (see Methods) (G^2 $p < 0.0001$, tau $= 0.49$ with $p = 0.0001$). When we control QMR type, the inverse relationship between temporality and QMR import value vanishes. Thus, we believe that temporal findings have smaller QMR import values only because they are more H&P-related, and therefore they do not carry the same weight as sophisticated tests. Figure 4 depicts a multiple-inheritance hierarchical classification of temporal types (i.e., more than one node in the abstraction tree can be the parent of a type) that captures their main features. A similar classification was developed for temporal reasoning abstractions (not shown here).

Frequency and importance of temporal entities

Of all 4,431 QMR findings, 17.5% were classified as temporal. Table 4 shows frequency distributions for some of the findings' attributes.

Table 4 ■

Frequency Distributions for Main Attributes

Among all findings:

TEMPORAL: yes 17.5%, no 82.5%

QMR TYPE: history 11.5%, symptom 5%, sign 25.4%, laboratory simple 6.5%, laboratory intermediate 30.7%, laboratory expensive/invasive 20.9%

IMPORT: low 2.3%, medium-low 15.8%, medium 35.5%, medium-high 32.4%, high 14%

Among temporal findings only:

TIME PRIMITIVES: implicit 93.3%, explicit points 0.3%, explicit intervals 6.4%

TIME UNITS: yes 5.5%, no 94.5%

TEMPORAL UNCERTAINTY: no 97%, yes 3%

PROCESSES: yes 45.9%, no 54.1%

REPEATING PATTERNS: yes 22.5%, no 77.5%

Table 5 ■

Temporal Patterns and Reasoning Types Found in Patients Cases, but Not in Quick Medical Reference (QMR)

Abstractions specifically pertaining to therapeutic planning

1. Multiple therapeutic changes until patient responds
2. Conditional temporal plans ("if X does not become Y within time period, I will do Z")
3. Intentions (admitted for . . . , therapy begun temporarily/permanently, etc.)

Abstractions that are interpretation of clinical actions and reasoning

4. Causal interpretations
5. Baseline value identification
6. Pending information
7. Patient orientation to time
8. Referral to unspecified time
9. History compatible/incompatible with

Abstractions that are elaborations of QMR abstractions

10. Often finding during episode
11. Seldom/from time to time/most of the time
12. Temporally qualified drug administration
13. Uncertain temporal progression
14. Nth episode out of M total episodes
15. Memory-related findings/tests/diseases
16. Generalities followed by exceptions (e.g., no history of X, besides a single episode)
17. Serial laboratory measurements
18. New vs old findings
19. Finding on and off during interval
20. Minimum/maximum values over an interval

Temporal reasoning found in medical records but not in QMR

We identified a number of temporal reasoning instances described in the medical record but not corresponding to a QMR temporal abstraction (Table 5). Most of those patterns and reasoning types are elaborations of existing QMR abstracted types, and constitute explanations of clinical actions and reasoning, or pertain to therapeutic plans. Thus, by not being specific to the diagnostic task, they would not, in our opinion, cause the system to have a decreased ability to derive a correct diagnosis.

Effects of Lack of Explicit Time on Diagnostic Performance

Most of the examined confounding variables were associated with only a small worsening of diagnostic performance (ORs between 0.54 and 0.74). Temporal

Table 6 ■

Odds Ratios (Confidence Intervals) for Diagnostic Accuracy, Given Various Case Properties

Temporal content: 0.70 (0.27–1.83)
Rareness: 0.71 (0.26–1.94)
Case length: 0.60 (0.23–1.57)
Uncertainty: 3.06 (1.15–8.13)
Spatial information: 2.36 (0.89–6.23)
Causal information: 0.63 (0.40–1.63)
Multiple diseases: 0.54 (0.05–6.25)
Deepness: 0.74 (0.25–2.18)

content had an OR of 0.7, which means that the odds (i.e., frequencies ratios) of getting a correct diagnosis vs an incorrect one in the high-temporal-complexity group was 70% the odds of a correct diagnosis vs an incorrect diagnosis in the low-temporal-complexity group. Unfortunately, our modest sample size did not allow for tight confidence intervals (95% CI = 0.27 to 1.83), and all the associations examined were not statistically significant (at the 0.05 level), so they must be interpreted as indicative only. Table 6 lists the ORs and CIs for temporal content and other explanatory variables. We used the Fastpro package[30] to derive high-density regions based on uniform priors for the probability of a successful diagnosis for both the high- and low-temporal-complexity content groups (assuming that ORs are log-normally distributed). In particular, the probability for an OR ≤ was 0.76, indicating that our data do not support strongly that temporal content adversely influences diagnostic performance. Figure 5 illustrates the high-density region for the OR of temporal content. Table 7 represents the cumulative distribution of this posterior distribution. For example, from Table 7 we see that an OR of ≥2 is true with a probability of ~0.01 given the data, while an OR of ≤0.5 is true with a probability of ~0.25. Since marginal independence does not necessarily imply conditional independence, we developed both logit[24] and Bayesian network multivariate models[29] for revealing a possible relationship between temporal content and diagnostic performance, conditioned on the variable context of the previously mentioned confounders. Figure 6 shows the probabilistic graphic model found to be the most probable in light of our experimental data, given the assumption that all models were equally likely a priori. In that model, diagnostic correctness is determined jointly by temporal content, number of diseases, uncertainty content, and spatial information. The model provides an interpretation of the dependency of diagnostic performance on temporal content and the rest of the variables in the form of a conditional probability distribution:

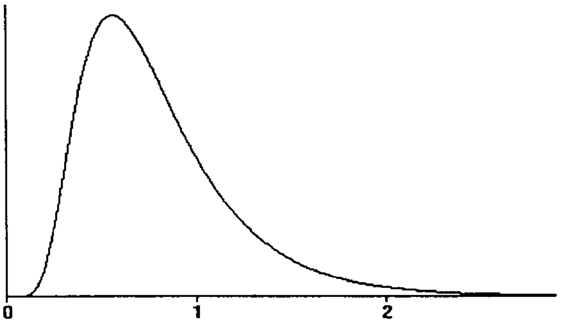

Figure 5 High-density region for odds ratio of temporal content.

$$p(\text{correct-diagnosis} \mid \text{temporal content, number of diseases, uncertainty content, spatial info})$$

By examining this distribution, we concluded that no clear form of covariation exists between temporal content and successful diagnosis, when the rest of the explanatory variables are taken into account. For instance, high temporal content is associated with low probability for correct diagnoses ($p = 0.17$) when the other three predictors take the value "high," while high temporal content is associated with high probability for a correct diagnosis ($p = 0.8$) when number of diseases is high and uncertainty is low. When we held the values of the confounding variables (number of diseases, uncertainty content, and spatial information) constant and observed the probabilities of successful diagnosis as a function of whether temporal content was low or high, sometimes the probability of a correct diagnosis increased, other times it decreased, depending on the set values of the three confounders. The use of measures of case content that were derived using principal components analysis did not yield any statistically significant predictors for diagnostic performance. Similarly, in logistic regression analyses, temporal content was not a statistically significant predictor for diagnostic performance.

Although the interpretation of these results is complicated and should be viewed with caution in light of the modest sample size, it suggests that *temporal content per se is not a strong indicator of the diagnostic performance of QMR.*

Discussion

In this paper, we reported an empirical analysis of QMR's implicit time both in terms of expressive power and diagnostic performance. We believe that work in medical AI has led to the accumulation over the years of epistemologically significant artifacts (e.g., MDSSs), the study of which can be of benefit to the development of a clearer understanding of important theoretical and engineering issues. We focused on one such system, and posed a particular question regarding the trade-offs between explicit and implicit time modeling. Although researchers have discussed circumstances when explicit time might not be necessary,[18] as well as why explicit time is important,[4,5,15,18] to our knowledge, there has not been an experiment specifically designed to clarify and study the effects of implicit modeling of time in medicine. Our results may appear counterintuitive because it seems easy to develop examples in which implicit time cannot deal appropriately with certain temporal reasoning queries. Additional arguments are based on case–control investigations of diagnostic failures. While existence proofs of cases where explicit time is indispensable suggest its necessity for at least some situations, they say very little about the heuristic power of carefully implemented MDSSs in highly complex problem-solving environments.[2] They also do not indicate how many cases that are currently diagnosed by atemporal systems would be incorrectly diagnosed (e.g., due to lack of explicit temporal knowledge) by some imperfect temporal implementation of the corresponding systems. Finally, they say little about the cost and benefit trade-offs between the two (radically different in terms of required development time and inferencing resources) approaches represented by implicit and explicit time modeling.

A concern about the present study has to do the relatively low post-hoc power (i.e., power estimated based on the observed effect) of many of the examined statistical tests. We believe that, although a high a priori power makes non-statistically significant results easier to interpret, one has to take into account the great cost of collecting patient cases and analyzing them at the level of detail we used, the fact that similar (or smaller) sample sizes have been used for important (non-temporal) evaluations of MDSS[7] (involving smaller effect sizes), and, finally, that our Bayesian analyses corroborate and complement the classic statistics conclusions. Moreover, recently researchers have criticized the use of pre-test power as arbitrary (since it depends on *arbitrarily large* estimates of the actual effect, and refers to a *class of outcomes* rather than a single outcome). Post-hoc power, on the other hand, seems to be uninformative as far as the interpretation of a statistical significance test (if the test is significant we do not care, but if it is non-significant it is *always low*). Post-hoc power is useful only for designing subsequent experiments based on our best information so far (i.e., the observed effect size). In light of those considerations, it has been proposed that CIs and Bayesian analyses are much more useful for the interpretation of non-significant results, an approach we adopted in this paper.[32]

Table 7 ■

Cumulative Distribution for Odds Ratio (OR) for Diagnostic Accuracy, Given Temporal Content

OR Threshold Value	p(True OR > OR Threshold Value)
0.20	0.995
0.50	0.750
0.70	0.500
0.98	0.250
1.82	0.050
2.18	0.010
3.15	0.001

Another interesting concern has to do with the quality of human abstraction and its effect in our study. Based on the experience of the human abstractors for our patient cases, we can claim that our study reflects a high-quality abstraction performance. Thus the conclusions are practically best-case in that respect.

An interesting improvement to this study would be to express the trade-offs in explicit vs implicit handling of time in decision theoretic terms rather than in terms of diagnostic accuracy. As it stands, our definition of correct diagnosis was, in simplistic terms, whether QMR included the primary GS in its full differential diagnosis (although we performed a sensitivity analysis on the size of QMR's diagnostic list, as described in section 2 of Methods). This is far from ideal in terms of *clinical significance*. Again, the prag-

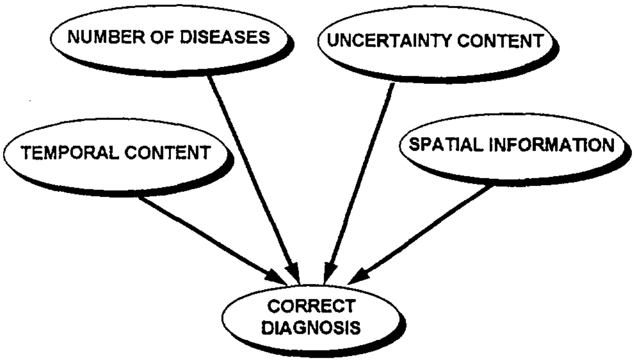

Figure 6 Determinants of diagnostic accuracy.

matics of designing and executing the evaluation reported here were the primary factors for this choice. Clinical experts suggested that it would be infeasible to come up with quality-adjusted life-year (QALY) estimates for patients at the level of complexity encountered in our cases, especially trying to take into account counterfactuals (e.g., QALYs given different diagnoses and corresponding treatments needed for assessing misclassification costs). Given these complexities, we believe that empirical analyses such as the one discussed in this paper are useful initial indications of the importance of abstracted time. Ideally, we would like to be able to study the problem of determining abstraction quality (with respect to some utility function characterizing the system's answers) in generalized terms and provide a theoretical treatment based on well-specified abstraction and utility function classes. Such an analysis seems extremely difficult to obtain, since it requires among other things the development of a formalism-independent means of expressing and analyzing the broad problem of medical temporal reasoning.

Summarizing our findings, at the knowledge-engineering level, we were surprised to find that the QMR KB contains an impressive array of different temporal types, which we identified and classified. The various temporal types are composed of a small number of primitives. We identified this ontology. We additionally abstracted specific temporal templates and types of temporal reasoning employed, and examined their importance. We believe that the identification of these temporal entities offers three potential benefits:

1. It *explains the ability of the system* to cope with the rich temporal nature of most patient cases, since it shows an abundance of temporal concepts that can be mapped to QMR findings. The ultimate utility of QMR's representation of temporal information is of course dependent on the human users of the system, who perform the abstraction from the patient record to the program. This study did not investigate that abstraction process. Also, in a few cases, the patient records were found to contain temporal statements about patients and their attributes that were not in the QMR lexicon.
2. In cases where a diagnostic system is designed to gather patient information without human abstraction, the study suggests the *types of temporal abstraction mechanisms (and thus intelligent temporal data pre-processors) that should be in place* for the system to function properly. These abstractions complement the set of suggested mechanisms offered by other researchers who have reported well-defined temporal abstraction mechanisms, aimed at having general applicability.[17,20]
3. In an exploratory sense, this study is a starting point for identifying *important temporal requirements for the design of formal MDSS models employing explicit time* (for example, among other things, it suggests that even explicit-time reasoners should allow the representation of abstracted time in order to exploit its heuristic value).

In the second part of the experiments described in this paper, we focused on diagnostic performance. We found that *temporal content has a modest, and statistically non-significant, effect on the diagnostic performance of QMR.* Although in the present study we demonstrated satisfactory heuristic power for the QMR system/domain with respect to the temporal robustness of its heuristic, implicit handling of time, we believe that only by obtaining a system-independent analysis of diagnostic performance with respect to temporal abstraction, we will be able to gain deeper insight into the limits of implicit time. It is evident that such an analysis is very difficult to obtain. It is important to reiterate that we do not argue against explicit time in MDSSs. Instead, we show experimental results that support the notion that temporal abstractions are a powerful heuristic for dealing with the intractabilities of explicit time. It is an open question *when* temporal abstractions can effectively substitute for explicit time, and *how* to develop them from fully or partially specified domain theories. Finally, it should be kept in mind that our findings are specific to the QMR system and domain. We hope that these results will stimulate similar analyses for other medical systems and domains, and that they will encourage MDSS developers to make judicious selections between abstracted and explicit time modeling.

The authors thank Drs. Michael Wagner, Johnathan Schwartz, Richard Ambrosino, and Sankharan Rajagopalan for their valuable input on the organization and presentation of the material that formed the basis of this paper. Dr. Henry Lowe provided helpful comments on earlier versions of the research design. Drs. Yuval Shahar, Peter Szolovits, William Long, and Harold Lehman provided challenging questions and suggestions that led to improvements of the discussion of the authors' conclusions.

References ■

1. Miller RA. Medical diagnostic decision support systems—past, present, and future. JAMIA. 1994;1:8–27.
2. Aliferis CF, Miller RA. On the heuristic nature of medical decision support systems. Methods Inf Med. 1995;34:5–14.

3. Allen JF, Hayes PJ. A common sense theory of time. Int Jt Conference Artif Intell proceedings. 1985:528–31.
4. Shoham Y. Temporal logics in AI: semantical and ontological considerations. Artif Intell. 1987;33:89–104.
5. Kahn MG. Modeling time in medical decision-support programs. Med Decis Making. 1991;11:249–64.
6. Miller RA, Pople HE, Myers JD. INTERNIST-I, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med. 1982;307:468–76.
7. Berner E, Webster GD, Shugerman AA, et al. Performance of four computer-based diagnostic systems. N Engl J Med. 1994;330: 1792–6.
8. Miller RA, Masarie FE, Myers JD. 'Quick Medical Reference' for diagnostic assistance. MD Comput. 1986;3:34–48.
9. Miller RA, McNeil MA, Challinor S, Masarie FE, Myers JD. Status report: The INTERNIST-1/Quick Medical Reference Project. West J Med. 1986;145:816–22.
10. Shortliffe E, Perreault L, eds. Medical Informatics: Computer Applications in Health Care. Reading, MA: Addison-Wesley, 1990.
11. Levesque HJ, Brachman RJ. A fundamental tradeoff in knowledge representation and reasoning. In: Levesque HJ, Brachman RJ, eds. Readings in Knowledge Representation. San Mateo, CA: Morgan Kauffman Publishers, 1985;42–70.
12. Stefanelli M. Therapy planning and monitoring [editorial]. Artif Intell Med. 1992;4:189–90.
13. Berzuini C, Bellazi R, Quaglini S, Spiegelhalter D. Bayesian networks for patient monitoring. Artif Intell Med. 1992;4:243–60.
14. Sager N. Medical Language Processing: Computer Management of Narrative Data. Reading, MA: Addison-Wesley, 1987.
15. Fagan LM. VM: Representing Time-dependent Relations in a Medical Setting [doctoral dissertation]. Stanford, CA: Stanford University, 1980.
16. Long W, Naimi S, Criscitielo M. Development of a knowledge base for diagnostic reasoning in cardiology. Comput Biomed Res. 1992;25:292–311.
17. Kohane I. Temporal reasoning in medical expert systems. MEDINFO 1986;170–4.
18. Keravnou ET, Washbrook J. A temporal reasoning framework used in the diagnosis of skeletal dysplasias. Artif Intell Med. 1990;2:239–65.
19. Das A, Tu S, Purcell G, Musen M. An extended SQL for temporal data management in clinical decision-support systems. SCAMC Proc. 1992;128–32.
20. Shahar Y, Musen MA. RESUME: a temporal-abstraction system for patient monitoring. Comput Biomed Res. 1993;26:255–73.
21. Elstein AS, Shulman LS, Spraska SA. Medical Problem Solving. Reasing, MA: Harvard University Press, 1978.
22. Ackerman E, Gatewood L. Mathematical Models in the Health Sciences. Minneapolis: University of Minnesota Press, 1979.
23. Chandrasekaran B, Wong T, Pryor T. 'Deep' models and their relation to diagnosis. Artif Intell Med. 1989;1:29–40.
24. Agresti A. Categorical Data Analysis. New York: Wiley Interscience, 1990.
25. Spector P. Research Designs. Thousand Oaks, CA: Sage Publications, 1981.
26. Parker RC, Miller RA. Creation of a knowledge base adequate for simulating patient cases: adding deep knowledge to the INTERNIST-1/QMR knowledge base. Methods Inf Med. 1989; 28:346–51.
27. Bankowitz R. The Effectiveness of QMR in Medical Decision Support. Report R01 HS06368.
28. Colton T. Statistics in Medicine. Boston: Little-Brown, 1974.
29. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Machine Learn. 1992;9:309–47.
30. Eddy DM, Hasselblad V. FAST*PRO: Software for Meta-analysis by the Confidence Profile Method. San Diego, CA: Academic Press, 1992.
31. Aliferis CF, Cooper GF, Buchanan BG, Miller RA, Bankowitz R, Giuse N. Temporal Reasoning Abstractions in QMR. University of Pittsburgh, Section of Medical Informatics, Technical Report SMI-94-03, 1994.
32. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med. 1994;121:200–6.