AI Magazine Volume 10 Number 1 (1989) (© AAAI)

A Computational Model of Reasoning from the Clinical Literature

Glenn D. Rennels, Edward H. Shortliffe, Frank E. Stockdale, and Perry L. Miller

This article explores the premise that a formalized representation of empirical studies can play a central role in computer-based decision support. The specific motivations underlying this research include the following propositions: (1) Reasoning from experimental evidence contained in the clinical literature is central to the decisions physicians make in patient care. (2) A computational model based on a declarative representation for published reports of clinical studies can drive a computer program that selectively tailors knowledge of the clinical literature as it is applied to a particular case. (3) The development of such a computational model is an important first step toward filling a void in computer-based decision support systems. Furthermore, the model can help us better understand the general principles of reasoning from experimental evidence both in medicine and other domains. Roundsman is a developmental computer system that draws on structured representations of the clinical literature to critique plans for the management of primary breast cancer. Roundsman is able to produce patient-specific analyses of breast cancer–management options based on the 24 clinical studies currently encoded in its knowledge base. The Roundsman system is a first step in exploring how the computer can help bring a critical analysis of the relevant literature, structured around a particular patient and treatment decision, to the physician.

© 1986 IEEE. Reprinted, with permission, from *Proceedings of the Tenth Annual Symposium on Computer Applications in Medical Care,* Washington, DC, October 25–26, 1986, pp. 373–380.

(See Editor's note at end of article)

AI research has increasingly emphasized the advantages of representing more fundamental knowledge about the problem domain than, for instance, a set of weighted links between observable findings and diagnostic hypotheses. Much of this work seeks to flesh out the causal models underlying diagnostic reasoning and represent these models (deep models) in an expert system to help drive its reasoning process. For example, an electronic circuit or a functioning human body is modeled, and computer programs are designed to search for causal explanations of malfunction (Patil, Szolovits, and Schwartz 1981; Davis 1984; Genesereth 1984). Planning medical management has not been as fully investigated, but several projects are currently exploring the notion that causal models of human pathophysiology can drive the analysis of medical management, for instance, by simulating the effects of perturbing homeostasis in different ways (Long et al. 1984).

When these models mirror a manufactured device (for example, an electronic circuit), causal models can indeed provide a sound basis for advice systems. In empirical sciences such as biomedicine, however, these models are secondary constructions, derived from experimental evidence. A medical example is breast cancer. Biological models of breast cancer are an unreliable basis for therapy planning, and the physician's reasoning must be directly grounded in the primary sources of experimental evidence (clinical trial publications).

Clinicians are well aware that good medical practice depends on keeping up to date with the clinical literature. To use this literature most effectively, a physician must critically assess these studies in the context of a particular patient and decide in what ways the experimental trial is relevant to the case at hand. Indeed, this very skill of recalling the key studies and evaluating how well the results apply to the patient is a process learned and practiced every day by teams of residents on their rounds. A computer system that fails to use this fundamental knowledge might, therefore, not fully capture the decision-making process of many medical domains. Nevertheless, little or no research has been conducted on the design of computer systems that reason explicitly from clinical studies to provide decision support for physicians.

This article describes a computer program in development, named Roundsman, that draws on structured representations of the clinical literature to critique plans for medical management. The goal of the Roundsman project is to model the process of reasoning from the clinical literature. Many medical domains exist in which such reasoning dominates. It is, therefore, important to explore how a machine might assist a clinician in this literature-based reasoning process.

The Roundsman project differs substantially from causal modeling in that there is no desire to model a device and its function but, rather, to model the structure of experimental evidence and its interpretation for decision making. In medical terms, it is not pathophysiological knowledge that is represented but knowledge about experimental trials and their relevance to a particular patient.

Nature of the Problem

How does a physician reason from the clinical literature? To gain insight into this question, the Roundsman project began with a period of informal protocol analysis. Previous work in medical protocol analysis investigated diagnostic reasoning (Kassirer, Kuipers, and Gorry 1982; Elstein, Shulman, and Sprafka 1978) and was particularly oriented toward causal models (Kuipers and Kassirer 1984). A senior oncologist at Stanford University Medical Center was asked to think aloud as he formulated management plans for primary breast cancer. These sessions were taped and later analyzed. By varying the clinical studies that the oncologist could draw on, it was possible to examine how a particular study contributed to the reasoning process and a study's role changed as additional studies were added to the library. We were particularly interested in how the oncologist's clinical judgment affected the interpretation of statistical results and the study as a whole.

Analysis of these transcripts suggested several organizational views. This section discusses one of these views (a publication-centered view) and its corresponding system design issues.

In the publication-centered view, critical readers embed the results of a clinical study in a matrix of contextual details that are attached to a particular clinical study. These contextual details help to interpret the meaning of the study's statistical results. These contextual details include the following: (1) What type of patients seek care at the hospital where the research was done? To use the study as a basis for treatment, a physician must assess the differences between the study population and a patient and decide whether these differences are likely to influence the outcome. (2) What is the track record of the author? Have previously published results been reproducible by other teams? (3) How qualified are the allied specialties that are involved in patient care but are not the subject of investigation, for example, postoperative nursing care? (4) What are the exact technical details for the treatments being compared (for example, two studies might compare the same drugs, but the dose and dosing schedules might differ)? Before the study can be used as a basis for therapy planning, a physician must consider whether the technical approach used in the study differed significantly from the approach that is planned. (5) How sophisticated is the biostatistical analysis?

An awareness of these contextual details allows physicians to decide how relevant the study is to their particular patient and treatment plan. The importance of this issue is evident in the examples provided and the design of a distance metric, as described later in this paper and in Rennels et al. (1986).

These contextual details overlay the study's experimental design and results. The design and results can have significant complexity themselves and frequently require analysis in assessing the study's relevance. For example, longitudinal, prospective comparisons of deliberate intervention (Bailar et al. 1984) compare one therapy group to another control group; these groups are optimally studied in parallel. Nevertheless, one of the most important sources of medical information has been the case series study, in which controls are external to the study (and, therefore, not formally matched at all).

Another dimension of design complexity concerns stratification. Patients are often sorted into strata according to variables thought to significantly influence their response to therapy. Results are then presented by stratum. Physicians can weed out many irrelevant tables and charts from the report if they can determine which stratum applies best to their patient. Even here, however, the critical reader exercises clinical judgment. For example, if the strata were constructed after treatment (post-stratification), one must assess the investigator's intent: Was the stratification motivated by genuine clinical concerns, or was it the product of a fishing expedition for a stratum that was statistically significant?

This publication-centered model, in which knowledge is structured around studies as distinct entities, also allows the natural representation of interstudy knowledge. For example, study A might have had an irregularity in the experimental design that left some doubt about the ability to generalize the main conclusions. Study B, published some time later, might demonstrate that the irregularity makes no difference, thus strengthening the principal conclusions of study A even though it might have investigated a different question.

One might structure this knowledge in other ways. The remainder of this section, however, focuses on four design goals of Roundsman's publication-centered implementation: (1) The computer system's data structures must reflect the publication-centered view. In particular, the system's critique of a treatment proposal for a particular patient must spring from declarative representations of one or more study's experimental design and observed outcome. (2) Convenient ways must exist to represent knowledge about the contextual details mentioned earlier, many of which are the subjective clinical judgments of our domain expert. (3) These contextual details must influence system performance in a substantive way. Ideally, the system should discuss the basic statistical results with the same priority and in the same manner as the domain expert. (4) The system must address clinical concerns in a realistic way and communicate in English well enough that clinical practitioners and biostatisticians can evaluate the potential of this approach to decision support.

The Roundsman System

The Roundsman program contains a library of clinical studies, each of which is represented as a separate data structure. These studies are not full-text copies of articles but high-level representations of the study's features. Examples of Roundsman using these studies to critique a physician's plan follow. Each example includes a verbatim transcript of Roundsman's current output.

To use Roundsman, the physician first describes the patient (for example, age 45, stage II breast cancer, premenopausal) and proposes a therapy choice (for example, surgical-wide excision [lumpectomy] followed by adjuvant radiotherapy). Roundsman produces a prose critique of the plan in light of the study. This critique is assembled dynamically and tailored to the particular patient, treatment decision, and clinical study(ies). The prose is generated by Roundsman's TEXTNET facility, an adaptation of Miller's (1984) PROSENET, which is based on the augmented transition net formalism.

In each example, Roundsman is critiquing a proposal for the surgical management of a particular patient's breast cancer. Because the critiques are rich in clinical detail, it must be reemphasized that Roundsman is currently a research project. These critiques are an important first step toward providing a new type of computer-based decision support, but they cannot be used as advice for actual clinical decisions at this time. Indeed, the trained clinician might notice comments that are clinically controversial or possibly incorrect. Further research and development are needed as well as additional intensive collaboration with the medical experts in the domain.

Each example consists of (1) a statement of the clinical context (a description of the patient and the physician's treatment proposal), (2) a verbatim transcript of Roundsman's critique of one particular study in light of this clinical context, and (3) a discussion of certain aspects of the example critique. For each clinical context, Roundsman usually selected between three and five studies for discussion. Ultimately, Roundsman will consider the overall impact of a group of studies. Currently, Roundsman critiques each relevant study serially. The output in these examples is typical of Roundsman's performance, and the only selection involved was ensuring that the examples included biomedical reports with different experimental designs, were from more than one time period, and showed how a study's critique changes for different clinical contexts.

Example 1

Clinical Context: For a woman whose tumor size is T1a, axillary lymph nodes are N1a, is premenopausal, age 45, unknown axillary node histology (cannot be known until after the surgery), and no metastases. Treatment proposal of surgical-wide excision (and axillary dissection) plus adjuvant radiotherapy plus adjuvant chemotherapy if axillary node histology is subsequently positive.

The following verbatim transcript shows Roundsman discussing the relevance of a publication by Fisher et al. (1985):

> Fisher85a was a randomized, controlled trial conducted at multiple NSABP centers. Subjects were assigned to verified wide excision with axillary dissection and adjuvant radiotherapy (N = 373) or the alternative therapy of verified wide excision with axillary dissection (N = 358). For patients who underwent the first protocol the proportion free of ipsilateral breast recurrence at five years was equal to 0.9, overall survival at five years turned out to be 0.91 and recurrence-free survival at five years was 0.81. Under the second protocol the proportion free of ipsilateral breast recurrence at five years turned out to be 0.77, overall survival at five years was 0.9 and recurrence-free survival at five years was equal to 0.68.
>
> Are these results relevant to your patient? It is encouraging that first, the adjuvant modality you propose was specified for this study as well (chemotherapy given if axillary nodes are path. positive). Second, this study population is quite similar to your patient (the women in this group had T sizes ranging up to T2a but excision margins were verified free of tumor). We suspect it makes little difference that the intervention was somewhat nonstandard (they did not radiate supraclavicular nodes). More troublesome is that the study population was probably in a better prognostic stratum than your patient (this study stratum was defined by negative axillary node histology; about 40% of clinical stage I patients like yours will have positive histology).
>
> What is the validity of the data? It helps that first, the investigator is reliable (the NSABP trials are first-rate, e.g., participating physicians must be certified by Fisher). Second, controls were randomly assigned. The results are weakened because one of their outcomes was a bit nonstandard (recurrence in the ipsilateral breast was NOT counted as a local recurrence).
>
> More than one outcome type is reported, but strictly on the basis of five-year results in recurrence-free survival, your suggested therapy seems best (although not all results agree). The close fit of your patient, considered together with the excellent methodology probably would not alter that statistical conclusion. The first protocol mentioned (which is close to your proposed plan) appears to be the better one.

Discussion of the Examples

Example 1 shows several reasons why the clinical literature is an interesting problem area for employing computer-based decision support.

Complexity of experimental design. This trial (NSABP protocol 6) compares three different interventions in parallel: (1) total mastectomy, (2) excision, and (3) excision plus radiation. In certain subsets of each group, chemotherapy was used. Several end points were reported: overall survival, recurrence-free survival, and ipsilateral breast recurrence. For the physician

Example 2

Clinical Context: For a woman with tumor size T1a, axillary lymph node status N1b, premenopausal, age 45, unknown axillary node histology, and no metastases. Treatment proposal of surgical-wide excision (and axillary dissection) plus adjuvant radiotherapy plus adjuvant chemotherapy if axillary node histology is subsequently positive.

The following is Roundsman's critique:

Fisher85a was a randomized, controlled trial performed by investigators at multiple NSABP centers. Patients were randomized to wide excision (and axillary dissection) and adjuvant radiotherapy and adjuvant chemotherapy (N = 229) or another protocol which was total mastectomy (& axillary dissection) and adjuvant chemotherapy (N = 224). For patients who underwent the first protocol the overall survival at five years turned out to be 0.75 and recurrence-free survival at five years was equal to 0.58. Under the second protocol the overall survival at five years was 0.66 and recurrence-free survival at five years was equal to 0.58.

How do these data apply to your patient? We are not particularly concerned that the intervention was somewhat nonstandard (they did not radiate supraclavicular nodes). More troublesome is that first, there were modifications to one intervention (in the excision arm, women with positive margins received total mastectomy, but remained in the "excision" group). Second, the study population was in a worse prognostic stratum compared to your patient (this study stratum was defined by positive axillary node histology; about 40% of clinical stage II patients like yours will have negative histology).

How much confidence can we have in the experimental results? It's good to see that first, the investigator is reliable (the NSABP trials are first-rate, e.g., participating physicians must be certified by Fisher). Second, controls were randomly assigned. The results are weakened because one of their outcomes was a bit nonstandard (recurrence in the ipsilateral breast was NOT counted as a local recurrence).

Looking selectively at five-year results in recurrence-free survival, those two interventions look equivalent (the other results generally agree). The "relevance" problems detailed above, considered together with the excellent methodology probably would not alter that statistical conclusion. Consequently, a choice between these two approaches might be made on the basis of morbidity (cosmesis, etc.) rather than cure.

requesting this consultation, Roundsman decided to highlight a comparison of intervention arms 2 and 3. Proper analysis of the results is complicated by the fact that the patients used to compute the results of intervention 2 are not the same when interventions 1 and 2 are compared and when interventions 3 and 2 are compared (as is discussed more fully in example 2). This design complexity has a domain-specific motivation that has more interest to oncologists than computer scientists. The important point is that even in presenting the first paragraph, Roundsman has already done a significant amount of work for the physician by sifting through the numerous interventions, subsets of patients, and end points in order to present selective portions of a complex body of experimental evidence. Next comes Roundsman's principal focus: the further subjective assessment of the relevance of these selective portions to the physician's clinical case.

Clinical details of the study. Although certain clinical details are crucial to an intelligent assessment of the study for clinical purposes, it is practically impossible for a physician to recall these details months or years after reading the article. The cost of refreshing these details is a line-by-line reading of a lengthy technical article. For example, certain clinical conditions had to exist before chemotherapy was given (paragraph 2); several T (tumor) sizes were allowed in this group of women studied, but excision margins were verified free of tumor (paragraph 2); supraclavicular nodes (that is, lymph nodes located above the collar bone) were not exposed to radiation (paragraph 2); and the definition of "local recurrence" excluded recurrences in the breast that had the original tumor (paragraph 3). Thus, Roundsman brings to light certain clinical details that might help the physician use this experimental evidence.

Relevance of clinical detail to the physician making a particular management decision. The second and third paragraphs not only offer subjective judgments about which clinical details of the study should be explicitly juxtaposed against the physician's patient and treatment decision but also offer subjective judgments about the importance of any mismatch when using the study to discuss the specific management problem. For example, the fact that the physician also plans to use chemotherapy if the axillary node histology turns out to be positive (paragraph 2) makes it easier to say this report can provide some support for this management decision. Irradiation of supraclavicular nodes is judged to be a minor detail whatever the physician chooses to do.

Example 2 shows how Roundsman's critique of Fisher (1985) (the study discussed in example 1) changes when the physician's patient is different.

In this example, Roundsman's critique has changed in several ways from that shown in example 1.

Complexity of experimental design. The patient in example 2 has a worse disease than the woman in example 1. This patient has stage II breast cancer, and it is more controversial whether excision is a safe surgical approach for her disease than for the woman in example 1. Consequently, in example 2, Roundsman chooses to focus its discussion on a comparison of a different surgical approach rather than a comparison of the omission or addition of radiation (as in example 1). As mentioned earlier, Fisher (1985) studied three different interventions in parallel; so, for this example Roundsman presented evidence concerning total mastectomy versus wide excision.

Relevance of clinical detail to the physician making a particular management decision. As mentioned in the discussion of example 1, the patients used to determine the results of intervention 2 are not the same when interventions 1 and 2 are compared as when interventions 3 and 2 are compared. Understanding this point requires attention to clinical detail: The Fisher (1985) protocol specified that women entered into the excision arm must have the margins of their excision verified free of tumor. If margins are not free, then the woman went on to have a total mastectomy. For the analysis of the excision group versus the total mastectomy group, women who failed to have clear margins (and thus received total mastectomies) were counted as members of the excision group. Why? To exclude them would have biased the results: The total mastectomy group did not check excision margins; excluding unclear margins from the excision group would exclude women with bigger tumors, making the results look better than they should. This clinical detail is brought to the attention of the physician in sentence 3 of paragraph 2 in the critique immediately previous.

One might then ask why this clinical detail was not mentioned in the critique in example 1. It was not because in comparing excision versus excision plus radiotherapy (discussed in example 1), women who fail to have clean margins and, therefore, receive total mastectomy are excluded from the count. The critique in example 1 need not concern the physician with clinical detail(s) of the study that do not affect the clinical context currently being considered.

Example 3 shows what a consultation with Roundsman would look like if the year was 1967. Thus, Roundsman is restricted to publications that appeared prior to 1968. Also, the date of the consultation (a system variable) is set to 1967. In 1967, total mastectomy was being advocated by some physicians, but it was a controversial management decision; the consensus was that any procedure less than a radical mastectomy endangered the life of a woman with breast cancer.

In examples 1 and 2, Roundsman discussed a randomized, controlled study. Example 3 shows Roundsman critiquing a clinical study that used nonrandomized internal controls. The publication discussed in this example (Peters 1967) appeared in the literature when radical mastectomy was the standard of care, and surgical excision was used by only a small minority of surgeons.

In retrospect, the nearly equivalent proportions reported in paragraph 1 have been borne out by later studies that compared mastectomy to excision (for example, Fisher [1985] in examples 1 and 2). In this consultation, however, Roundsman is unable to confidently conclude that Peters (1967) provides enough support for the physician to deviate from more standard surgical approach (see the last sentence of the concluding paragraph). The reasons for this lack of support are explained by Roundsman: the nonuniform nature of the intervention (paragraph 2), the broad stratum of patients lumped together for analysis (paragraph 2), the nonrandom experimental design (paragraph 3), and the long accrual period (paragraph 3).

System Design of Roundsman

This section provides an overview of the Roundsman system's flow of control and describes Roundsman's principal data structures.

Example 3

Clinical Context: For a woman whose clinical exam reveals tumor size T1a, axillary nodes thought to contain tumor (N1b), is premenopausal, age 45, and has no metastases. Treatment proposal of wide excision, axillary dissection, and adjuvant radiotherapy.

What follows is Roundsman's critique:

Peters67 employed non-randomized controls in a study conducted at Princess Margaret Hospital, Toronto. A set of patients who were treated by wedge resection and adjuvant radiotherapy (N = 94) is contrasted to a second therapy group: radical mastectomy and adjuvant radiotherapy (N = 247). In the group which received the first protocol the overall survival at five years was 0.76. For patients who underwent the second protocol the overall survival at five years turned out to be 0.72.

How well does the study generalize to your particular patient? We are not particularly concerned that one modality you propose may not be quite like what was done in the study ("wedge resection" here indicates excisional biopsy, quadrant resection, or any technique to excise the primary). More troublesome is that the study population was probably in a better prognostic stratum than your patient (the study used a pooled clinical stage I and II—so that's a slightly better prognostic group than your patient).

How much confidence can we have in the experimental results? The results are weakened because first, choice of treatment was decided nonrandomly (nor were subjects and controls matched on prognostic parameters). Second, patients were accrued over a rather long period (this is a retrospective study of patients treated between 1935 and 1960). Third, this is a wide stratum (it would have been preferable to separate stages I and II).

Considering the reported observations and sample size (see introductory paragraph), those two interventions look equivalent. The small mismatch of your particular clinical situation, considered together with the large methodological weaknesses however, leads us to think that the results are indecisive for your purposes. Adhering to the standard of care (radical mastectomy) would probably be most appropriate.

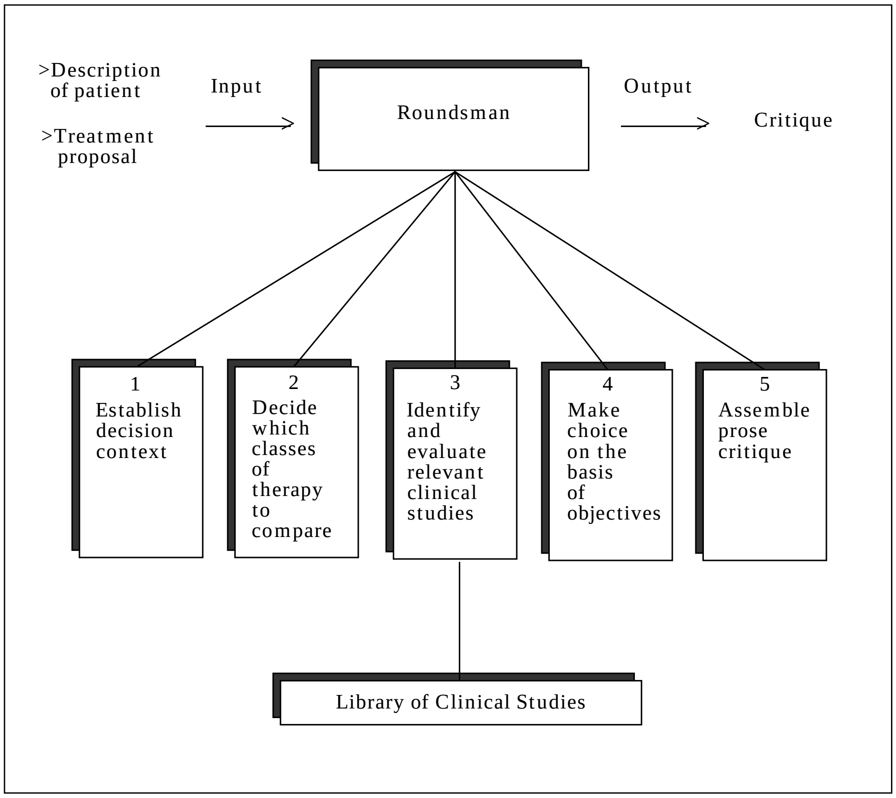
>Description of patient
>Treatment proposal
Input
Roundsman
Output
Critique
1 Establish decision context
2 Decide which classes of therapy to compare
3 Identify and evaluate relevant clinical studies
4 Make choice on the basis of objectives
5 Assemble prose critique
Library of Clinical Studies

Figure 1. Flow of Control in the Roundsman System.

Control

The steps taken by Roundsman when analyzing a case (see figure 1) follow:

1. Establish the *decision context.* The decision context includes information about the patient and the therapy that the physician is proposing for this patient.

2. Focus on the class of questions most likely to interest the physician. This step entails deciding what types of therapeutic intervention should be compared. For example, in one time period, it might be more appropriate for the machine to first discuss the surgical procedure, whereas in another time period, it would be more appropriate to first discuss the use of, or omission of, adjuvant radiation with the proposed surgery. The need to establish an appropriate focus results because the clinical consensus changes over time, as discussed in detail later.

3. Determine for each study in the library whether it can provide experimental results concerning the class of questions. If so, then first find the group (stratum) of patients within the study that most closely approximates the physician's patient. Second, identify any experimental results of the stratum that was treated with the interventions of interest (see step 2). Third, assess the "distances" between the physician's decision context and the particulars of the clinical study. (This process is discussed further when Roundsman's distance-metric and distance-estimator are described.) Fourth, return the study results as applied to the chosen stratum, together with the distance assessments, to higher-level control functions in Roundsman. All this information is packaged in an object called a *datum-from-study.*

4. Use the datum-from-study to compare alternative interventions on the basis of a model of choice and explanation.

5. Pass the conclusions of the system to a prose generation module that assembles a prose critique for the user.

Data Structures

The Roundsman system is organized around frame-based data structures. The most prominent of these data structures is the study. For example, Peters (1967) is 1 of 24 publications currently represented in Roundsman's library. The heart of each study consists of strata (sets of the data structure "stratum") and comparisons (sets of the data structure "comparison of interventions"). In addition, each study contains certain descriptive information, such as the name of the institution where the research was conducted.

A *stratum* is a definition of the study population. Publications of clinical studies do not report data at the level of individual patients. Thus, a stratum is not a collection of patients as in a data base but is a summary description for a population. Consequently, the central component of each stratum is a population-description, a data structure (not text strings) that can be interpreted by the machine. For example, one population-description follows:

```
a POPULATION-DESCRIPTION with
   clinical-stage-set = (I II)
   t-set = (T0 T1a T1b T2a T2b)
   n-set = (N0 N1a)
   path-n-set = (UNKNOWN)
   m-set = (M0)
   menopausal-status-set = (PRE POST)
   age-lower-bound = 20
   age-upper-bound = 80
```

This population-description includes patients of varying tumor sizes (T0 through T2b), a narrow range of clinical node status (N0 and N1a), unknown axillary node pathology, no distant metastases (M0), and a wide age span.

In addition to strata, a key part of any study is the *comparison object* (comparison of interventions). Each comparison contains knowledge about an experiment comparing one therapeutic intervention with another. For example, one comparison from Peters (1967) follows:

```
a COMPARISON with
   study-id = Peters67
   stratum-id = 1
   comparison-id = 1
   intervention-A = <pointer to an inter-
   vention>
   intervention-B = <pointer to an inter-
   vention>
   sample-size-A = 203
   sample-size-B = 609
```

```
outcome-A = <pointer to an outcome>
outcome-B = <pointer to an outcome>
standard-error-of-difference = 0.056
```

This comparison encodes details about the interventions being compared, the stratum involved, and the results that were measured. In addition, each comparison might identify the study it belongs to (id Peters67) and other comparisons that compare the same two interventions but concern different results (id's 2, 3, and 4). For example, the outcome of the comparison shown here (with id 1) might pertain to overall survival at 5 years, whereas comparison 2 might pertain to recurrence-free survival at 10 years. The motivation for separating this information into separate comparisons is not discussed here. Each of these components is, in turn, a data structure. For example, Roundsman has an outcome hierarchy in which "5 year survival" is one "measure of overall survival."

If one could meaningfully reason on the basis of statistical grounds alone, almost all the study's information could be captured by knowing the type of patients, the sample size, the two interventions, and the results. As discussed earlier, however, these studies cannot be used productively in clinical reasoning without expert clinical interpretation. Roundsman would provide little of value if it merely offered the statistical skeleton. Consequently, each comparison also possesses *distance metric knowledge,* which is used to evaluate the clinical relevance of the statistical results to a particular patient and treatment.

In its operation, Roundsman selects certain comparisons (data structures located inside studies), each of which is augmented by its own dynamically assessed distance metric (Rennels et al.1986). This metric consists of a set of distances assessing how well the comparison applies to the patient and the proposed plan. These distances include population mismatches, intervention mismatches, and methodological weaknesses. Two examples of metric components follow:

```
a LONG-ACCRUAL-PERIOD with
  se-change = INCREASE-SMALL
  dp-change = NONE
  specifics = "patient entry lasted from
  1939 to 1972."

a BETTER-PROGNOSTIC-STRATUM
  dp-change = AWAY-FROM-ZERO-
  SMALL
  specifics = "the study used a pooled
  clinical stage I and II—so that's a
  slightly better prognostic group than
  your patient"
```

The labels dp-change and se-change are slot titles referring to effects of this metric component on the difference between proportions (dp) of intervention-A and intervention-B and the standard error (se) of this difference.

Components of this distance metric are used by Roundsman to analyze the clinical and statistical relevance of a particular comparison to the problem at hand. For example, in order to generate the prose output shown in the examples, the metric components were first divided (dynamically) according to whether they were mismatches with the particular patient and treatment proposal or methodological issues. Within the first group, components were further divided into three subgroups: good matches, mismatches that are negligible in overall impact, and mismatches which are significant. Similarly, methodological issues were sorted into good methodology, methodological weaknesses of negligible impact, and serious weaknesses. Roundsman then assembled a prose discussion in the context of these subdivisions.

The metric knowledge associated with a comparison consists of one or more distance-estimators. Each distance-estimator contains clinical heuristics and judgments collected from our oncologist domain expert. Distance-estimators are capable of contributing to (and, thus, enlarging) the distance metric associated with a comparison. For example, the distance-estimator shown here would insert a "better prognostic stratum" distance component into the distance metric if, for the proposed treatment, a study population is in a better prognostic stratum than the physician's patient.

```
a POPULATION-DISTANCE-ESTIMA-
TOR with
  outcome-eq-classes = (OAS)
  intervention1-eq-classes = (ANY)
  intervention2-eq-classes = (ANY)
  study-pop-classes = (CLINICAL-
STAGES-I-II)
  patient-classes=(CLINICAL-STAGE-II)
  bias-incurred=(a BETTER-PROGNOS-
  TIC-STRATUM with dp-change=
  AWAY-FROM-ZERO-SMALL
  specifics = "the study used a pooled
  clinical stage I and II - so that's a
  slightly better prognostic group than
  your patient" )
```

The distance-estimator lists equivalence classes, which are defined on the system's results, interventions, population-descriptions, and patient-descriptions. The system has population distance-estimators (to assess mismatches between a study population and a patient) and intervention distance-estimators. The population distance-estimator shown here is activated if (1) the outcome being discussed is a member of the OAS equivalence class (any measure of overall survival), (2) the interventions being considered are any of the interventions known to the system, (3) the study stratum being examined by Roundsman is composed of subjects who were clinical stage I or stage II, and (4) the physician's patient is clinical stage II. The result of activating this distance-estimator is the insertion of a better prognostic stratum distance into that comparison's metric.

If Roundsman is applied to problems other than breast cancer, certain data structures must be changed (for example, results and interventions). However, much of the current implementation can be generalized to apply to clinical studies in other medical domains.

Research Contribution

The research contribution of this project can best be understood by viewing Roundsman from the perspectives of AI, medical decision analysis, and bibliographic retrieval.

AI techniques are being applied to an increasing variety of problems. Biomedicine and the social sciences repeatedly present problem domains for which no reliable causal models exist. In these domains, system designers might retreat to the surface-level heuristics that sufficed for first-generation expert systems. Instead, we suggest the investigation of how experts reason from the relevant bodies of experimental evidence. This evidence might well have its own structure (as is the case for clinical literature), which is tremendously useful when combined with knowledge about how to reason based on this structure. Building computer-based models of this reasoning process might yield useful decision support systems, and it might illuminate general principles of reasoning from experimental evidence, opening these principles to further explicit analysis.

One of the most difficult and time-consuming parts of performing medical decision analysis is estimating the probability of events. It is a task that requires a strong clinical background and experience reading biostatistical reports. Furthermore, this task is common to a variety of methodological approaches, from standard decision trees to Markov processes. Little explicit analysis has been conducted, however, of the reasoning process by which probabilities are assigned and (to our knowledge) no attempts to model it in a computer-based advice system. The Roundsman project explores the underlying reasoning process involved in making these assessments.

Unlike many current computer-based medical advice programs, bibliographic retrieval systems often meet immediate enthusiasm by clinicians. In these systems, full-text copies (or abstracts) of journal articles are retrieved by a keyword index, which can be organized in a disease hierarchy or according to the keyword's proximity to another keyword. These journal articles have the potential to change management decisions (Scura and Davidoff 1981). The state of this science, however, is quite primitive: Matching strings of alphanumeric characters falls far short of intelligent information retrieval. The current Roundsman system is a step toward the development of systems that understand the structure of the literature they are searching and can make inferences about how an article might relate to the clinical problem a physician faces.

Summary

The Roundsman project contributes to the better understanding and development of fundamental models of medical decision making. The approach differs substantially from causal modeling in that there is no desire to model human pathophysiology but rather to model the structure of experimental trials and their relevance to a physician's patient and treatment plan. The development of this computational model suggests a promising new direction for medical informatics; decision support systems that bring a critical analysis of the relevant literature structured around a particular patient and treatment plan to the physician might be a vital addition to the tools of practicing physicians. Furthermore, computational models of how physicians reason from the clinical literature can illuminate general principles of reasoning from experimental evidence, opening these principles to further explicit analysis.

Acknowledgments

This paper also appears in the Proceedings of the Tenth Annual Society for Computer Applications in Medical Care (Rennels 1986). Rennels was supported by training grant LM-07033 from the National Library of Medicine (NLM). Shortliffe is a Henry J. Kaiser Family Foundation scholar in general internal medicine. Miller is supported in part by NLM grant LM-04336. Computing resources were provided by the Biomedical Research Technology Program under grant RR-00785 and the Hewlett-Packard Company.

References

Bailar, J. C.; Louis, T. A.; Lavori, P. W.; and Polansky, M. 1984. A Classification for Biomedical Research Reports. *New England Journal of Medicine* 311:1482–1487.

Davis, R. 1984. Diagnostic Reasoning Based on Structure and Behavior. *Artificial Intelligence* 24: 347–410.

Elstein, A. S.; Shulman, L. S.; and Sprafka, S. A. 1978. *Medical Problem Solving: An Analysis of Clinical Reasoning.* Cambridge, Mass.: Harvard University Press.

Fisher, B.; Bauer, M.; Margolese, R. et al. 1985. Five-Year Results of a Randomized Clinical Trial Comparing Total Mastectomy and Segmental Mastectomy with or without Radiation in the Treatment of Breast Cancer. *New England Journal of Medicine* 312: 665–673.

Genesereth, M. R. 1984. The Use of Design Descriptions in Automated Diagnosis. *Artificial Intelligence* 24: 411–436.

Kassirer, J. P.; Kuipers, B. J.; and Gorry, G.A. 1982. Toward a Theory of Clinical Expertise. *American Journal of Medicine* 73: 251–259.

Kuipers, B. J., and Kassirer, J. P. 1984. Causal Reasoning in Medicine: Analysis of a Protocol. *Cognitive Science* 8: 363–385.

Long, W. J.; Naimi, S.; Criscitiello, M. G.; Pauker, S. G.; and Szolovits, P. 1984. An Aid to Physiological Reasoning in the Management of Cardiovascular Disease. In Proceedings of the IEEE Computers in Cardiology Conference, 3–6. Washington, D.C.: IEEE Computer Society Press.

Miller, P. L. 1984. A *Critiquing Approach to Expert Computer Advice: ATTENDING.* Boston: Pitman.

Patil, R. S.; Szolovits, P.; and Schwartz, W. B. 1981. Causal Understanding of Patient Illness in Medical Diagnosis. In Proceedings of the Seventh International Joint Conference on Artificial Intelligence, 893–899. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence.

Peters, M. V. 1967. Wedge Resection and Irradiation. *Annals of Internal Medicine* 200: 144–153.

Rennels, G. D. 1986. A Computational Model of Reasoning from the Clinical Literature. In Proceedings of the Tenth Annual Symposium on Computer Applications in Medical Care, 373–380. Washington, D.C.: IEEE Computer Society Press.

Rennels, G. D.; Shortliffe, E. H.; Stockdale, F. E.; and Miller, P. L. 1986. Reasoning from the Clinical Literature: A "Distance" Metric. In Proceedings of the 1986 American Association for Medical Systems and Informatics Congress, 19–23. San Francisco, Calif.: American Association for Medical Systems and Informatics.

Scura, G., and Davidoff, F. 1981. Case-Related Use of the Medical Literature: Clinical Librarian Services for Improving Patient Care. *Journal of the American Medical Association* 245: 50–52.

Editor's Note

This article first appeared in the Proceedings of the Symposium on Computer Applications in Medical Care (SCAMC). The symposium took place in October 1986. The paper won first prize in the student paper competition sponsored by SCAMC and the research that is described in this paper received a national award in 1988. Although we do not ordinarily republish articles that have appeared in proceedings or other periodicals, we are making an exception in this case because we believe it deserves wider distribution.

Glenn D. Rennels is a graduate of the Medical Information Sciences Program at Stanford and just completed a residency in anesthesiology at the Stanford University School of Medicine. His current address is 2115 Princeton St., Palo Alto, CA 94306. His is interested in computer science research that has potential for medical applications.

Edward H. Shortliffe is associate professor of Medicine and Computer Science at Stanford University. He is the scientific director of the Medical Computer Science Group in the Knowledge Systems Laboratory and chief of the Division of General Internal Medicine at Stanford University School of Medicine. His research interests include models for evidential reasoning and representation techniques to support advanced explanation capabilities for expert systems.

Frank E. Stockdale is a professor of medicine in the Division of Medical Oncology at the Stanford University School of Medicine, Stanford, CA 94305. His expertise is in the clinical management of breast cancer.

Perry L. Miller is an associate professor in the Department of Anesthesiology and the director of the Medical Informatics Program, Yale University School of Medicine, 333 Cedar Street, P.O. Box 3333, New Haven, CT 06510. His research interests include AI applications in medicine and other topics in medical informatics.