Use of Linear Models to Analyze Physicians' Decisions

ROBERT S. WIGTON, MD

Linear models of judgment are powerful tools for studying medical decision making. The recent increase in applications of these models to medicine reflects more available computing resources and the parallel development of clinical prediction rules derived from multivariate analysis of patient data. Psychological research into expert and novice decision making shows that linear models derived from judges' decisions usually predict future decisions more accurately than either the judge or a mechanical application of the judge's stated policies. Studies of medical decision making have shown similar results, as well as marked variation among experts in how they appear to use clinical information. Cognitive feedback, which is feedback to the learner of the judgment model derived from previous decisions, is highly effective for teaching complex judgment tasks. Many technical problems remain to be mastered in constructing linear models of medical judgment. These include how to select the correct variables, how to provide a selection of variables broad enough to accommodate individual variations in strategy, how to model intercorrelated variables, and how to characterize and aggregate individual strategies. Despite the methodologic challenges, linear models remain a powerful method for studying how physicians combine multiple items of imperfect information to make a judgment. These techniques may provide important insights into variation in physician judgments. In addition, they hold promise in teaching the appropriate integration of complex data in the day-to-day practice of medicine. *Key words:* linear models; medical judgment; diagnosis; prognosis. **(Med Decis Making 8:241–252, 1988)**

Linear models are particularly well suited to studying the complex judgments involved in medical diagnosis and treatment. Their computational intensity limited early development and application, but the widespread availability of computing resources has stimulated an increasing interest in linear models.

Linear models portray judgment as the sum of important factors either for or against a decision or diagnosis multiplied by the relative importance (weight) of each factor. A judge's strategy can be inferred from decisions made over a series of cases where the status of these important factors (cues) is known. For example, one can calculate the weight a physician places on the presence or absence of chest pain in diagnosing myocardial infarction by observing how his or her diagnosis changes as the presence of chest pain varies over a large sample of patients (while controlling for other important diagnostic features). This method often is called "policy capturing" or, more recently, "judgment analysis."

Linear models are well suited to investigating medical judgment because their structure matches the way these tasks have been viewed from within medicine: the combining of clinical findings of uncertain predictive power to arrive at a diagnosis or management plan. Clinical prediction rules, linear models derived from multivariate analysis of clinical data, are becoming more prevalent in clinical medicine.[60] In addition, less than optimal judgment or wide variation in judgments is of great importance in medicine because of the effects on patient welfare and the cost of care.

After describing some early studies in the development of judgment analysis, I focus on how these techniques have been used to study decisions regarding medical diagnosis and therapy. I then describe specific methods that have evolved to construct and analyze simulated clinical cases, with comments on the advantages and disadvantages of each. Last, I discuss future directions and challenges. This review focuses on the use of linear models for capturing judgment strategies, and I do not review research using linear models to determine utilities and values for decision making. Readers interested in this area are referred to Von Winterfeldt and Edwards.[58]

Early Use of Linear Models in Decision Making

The idea of using an additive linear rule to aid in decision making is not new. Dawes and Corrigan,[7] in their review of linear models, cite Benjamin Franklin's recommending the use of such a model in a letter to Joseph Priestly in 1772. Thorndike,[55] in a 1918 article entitled "Fundamental Theorems in Judging Men," proposed a weighted linear model to judge applicants

Received September 28, 1987, from the Section of General Medicine, Department of Internal Medicine, University of Nebraska College of Medicine, Omaha, Nebraska. Supported in part by grant 1RO1 LM 04321 from the National Library of Medicine. Second of three papers from the Advanced Short Course, Eighth Annual Meeting of the Society for Medical Decision Making, Chicago, October 1986.

Table 1 • Wallace's Comparison of the Weights Used by the Corn Judges with the Weights Derived from the Actual Yield[59]

	Relative Weight	
	From Corn Judges' Estimates	From Actual Yield
Length	42.0	7.7
Circumference	13.6	10.0
Weight of kernel	18.3	50.0
Filling at tip	13.3	18.0
Blistering	6.4	9.0
Starchiness	6.4	5.3
	100.0	100.0

for any position where the desired characteristics are known. In a study that anticipated some of the later research, Henry Wallace, an agronomist and later Secretary of Agriculture and Vice President under Roosevelt, studied the weights that experienced corn judges used to predict future yield after inspecting individual ears of corn.[59] First, Wallace calculated the correlations between the characteristics of over 1,500 ears of corn and the yield the corn judges predicted for each ear. Next, he determined the correlation between the characteristics of each ear and the yield, which he knew from actual measurement. He found that the weights the corn judges gave the individual characteristics differed from the way those characteristics related to actual yield (table 1). In addition, the judges were not at all accurate in predicting yield. Wallace concluded that the judges' lack of accuracy in predicting yield related to inappropriate weighting of the characteristics of the ears of corn. The low correlation he found between the judges' weights and the optimal weights would be a prominent finding in later research.

Brunswik's Lens Model

Egon Brunswik gave the linear model of judgment its most elegant conceptual form.[19] Brunswik used the analogy of rays of light passing through a convex lens to describe his concept of the relationship between the interpretation of information (cues) and the actual relationship of those cues to the real world.

Hammond and others have developed and enhanced Brunwik's "lens model" into a general method for studying intuitive judgments[17,18,22,23,27,56](fig. 1). This method uses multiple linear regression to calculate the weight for each cue and uses the percentage of variance explained by the regression model of the judge as a measure of the judge's consistency. The correlation between judgments made and the true outcome, when known, provides a measure of accuracy.

Studies using this model have repeatedly shown considerable variation among expert decision makers in how they used the available information to reach a judgment. In addition, when trying to predict future decisions, investigators found that the weights derived from the judges' answers often outperformed the weights each judge thought he or she was using.[6,13]

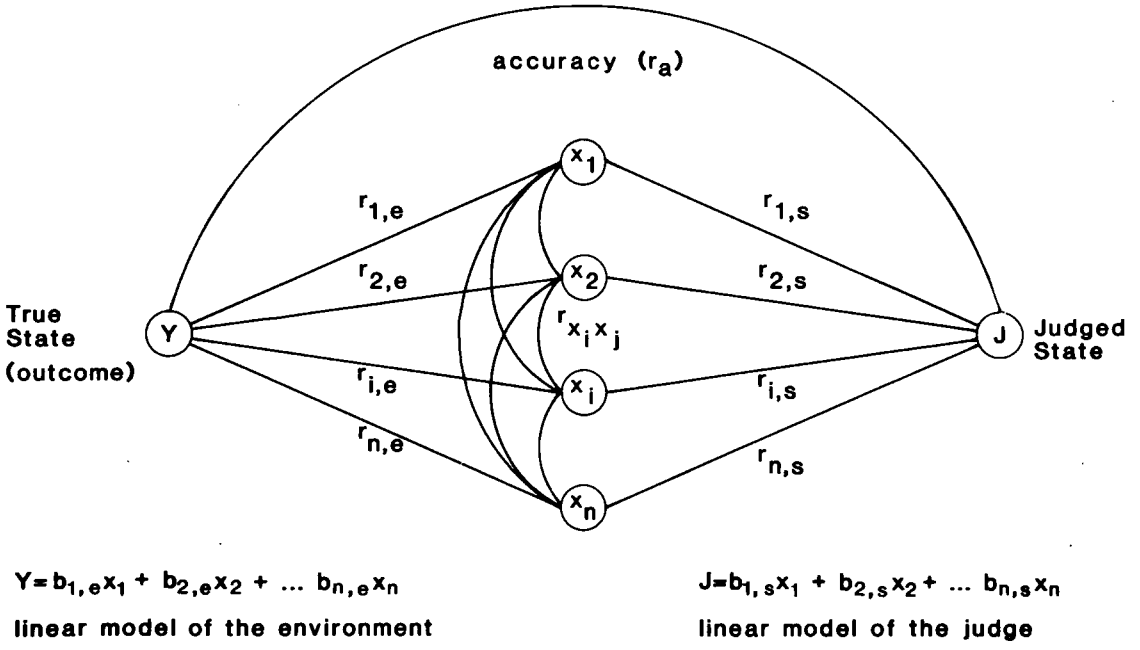

FIGURE 1. The lens model. A linear model of the judge is calculated from the repeated judgments on the right side of the lens while a model of the environment is calculated from repeated observations of actual outcomes, on the left side. $x_1 \ldots x_n$ represent the cues. $r_{1,s} \ldots r_{n,s}$ are the correlations between the individual cues and the judgments made and $r_{1,e} \ldots r_{n,e}$ are the correlations between the individual cues and the actual outcomes (ecology). $r_{x_i x_j}$ represents the interactions between cues. Accuracy is represented by the correlation between the judgments and the actual outcomes.

Table 2 • Use of Linear Models of Judgment in Medicine

Type of Problem	Topic (Reference)	Source of Model*
Diagnosis or prognosis	Patient response to anesthesia[18]	PC
	Radiologic diagnosis of gastric cancer[49]	PC
	Prognosis from pathology in Hodgkin's disease[8]	PCA
	Seveity of depression[9]	PC
	Clinical diagnosis of pulmonary embolism[61]	PC
	Disease activity in rheumatoid arthritis[30,31,32]	PCA, PC
	Diagnosis of hyperactive children[57]	PCA, PC
	Clinical diagnosis of streptococcal pharyngitis[39,61]	AC, PC
	Patient responses to drugs[29,53,54]	PC
Patient management decisions	Use of psychoactive drugs[10,12]	PC
	Patient compliance with antihypertensive drugs[44]	PC
	Referral of obese patients[45,46]	PC
	Referral of high-risk obstetric patients[42]	PC
	Tube feeding in seriously ill patients[51,52]	PC
	Testing for hypertensive patients[46]	PC
Assessment of trainees	Performance of anesthesiology residents[36]	PC
	Candidates for surgical residency[5]	PC
	Candidates for internal medicine residency[64]	AC
Cognitive feedback	Diagnosis of urinary tract infection[61]	PC
	Diagnosis of streptococcal pharyngitis[63]	PC

*PC = "paper" (simulated) cases; PCA = "paper" cases from actual cases; AC = actual cases.

Application of Linear Models of Judgment to Medicine

In the following sections I review specific applications of judgment analysis to medical problems. These studies (Table 2) illustrate applications to problems in medical diagnosis, patient management, trainee assessment, and ethical decisions. Most studies have used "paper" cases with limited numbers of variables to model physicians' strategies, but in a few cases, the strategies have been modeled from observation of clinicians dealing with actual cases.

In these studies, the findings of most practical importance to medicine have confirmed findings from judgment analysis in other areas. The medical studies have found wide variation among medical experts in how they use available information to reach decisions. In predicting future decisions, linear models of medical experts' strategies often outperform the experts' self-reported strategies. In addition, several medical studies have found variables given considerable weight by physicians in making judgments about "paper" cases, but not considered important by the physicians themselves.

Some early investigations of clinical judgment featured medical problems, such as Hammond's and Hering's 1955 studies of anesthesiologists' policies for predicting patient responses to anesthesia,[17] but the earliest report in a medical journal appeared many years later.[49] In this study, published in 1971, Slovic and colleagues modeled radiologists' diagnosis of gastric cancer by asking expert radiologists to judge whether 24 simulated cases represented benign or malignant gastric ulcers based on the presence or absence of seven radiologic signs.[49] As in previous studies of non-medical tasks, the authors found great diversity among expert radiologists in the weights they gave each finding in making a judgment. At about the same time, Einhorn was studying the strategies of four physicians as they used nine biopsy characteristics to forecast the outcomes of 200 patients with Hodgkins' disease.[8] Although physicians' overall rating of severity did not correlate with survival time, the accuracy of predictions could be enhanced by incorporating the physicians' observations in a linear equation.

MODELING MEDICAL DIAGNOSIS OR PROGNOSIS

Despite these early studies' success in using linear models to study diagnosis, they and linear modeling research in particular have only recently become known within the medical community. Studies of medical diagnosis have continued to find a surprising diversity of judgment among medical experts.

Fisch and colleagues investigated how physicians diagnose and manage depressed patients by asking 15 general physicians to judge the severity of depression in 80 case simulations.[9] They found, as had Slovic,[49] poor agreement among physicians in how to use the clinical information to make such judgments.

We found similar variation in diagnostic strategies when we examined how experienced physicians and students diagnose pulmonary embolism from simulated cases.[62] Although we had predicted that diagnostic strategies would become more similar as clinical

experience increased, we found instead that the strategies of experienced faculty were as diverse as those of junior medical students and house officers. Furthermore, diagnostic weighting differed in several respects from optimal weights derived from actual cases.

Kirwan and colleagues studied how rheumatologists weighted patients' clinical signs and symptoms to judge disease activity in rheumatoid arthritis.[30-32] They, too, found great variation in physicians' diagnostic strategies as determined from their decisions about "paper" cases.[30] Although inconsistency in judgment played some role, much of the variation among physicians in this and the studies cited above reflected differences in the judgment policies (weights) themselves.[31] In a subsequent study, Kirwan found that physicians' weights, derived from their responses to "paper" cases, predicted their responses to new cases better than either a strategy of equal weighting or the strategy the physicians thought they were following.[33] Ullman and Doherty found similar variation in experts' strategies. They investigated how physicians diagnose hyperactivity in children by asking them to review "paper" cases derived from clinical records.[57] They concluded that a major determinant of whether a child was diagnosed as hyperactive was who was making the diagnosis.

Rather than use paper cases, Poses and colleagues studied how student health physicians diagnosed streptococcal pharyngitis by calculating their weightings of clinical signs and symptoms from diagnostic estimates made on actual patients presenting with sore throat.[38] They not only found poor agreement among the physicians but also found that their probability estimates were inaccurate[39] (also, Poses RM, personal communication).

In the 1970s, Joyce, Stewart, and colleagues began a series of studies that modeled physicians' strategies judging a patient's clinical response to therapy in research trials.[53] By applying a model of the physicians' judgment policy to the various indicators of the patients' clinical states, they were able to reduce experimental variation due to differences in judgment and thus reduce the number of cases needed to achieve significance in the drug trials. This line of research has resulted in recent proposals for improving the efficiency and power of clinical trials.[29,54]

MODELING CLINICAL DECISIONS

Several investigators have used simulated clinical cases to model physician management decisions. Gillis et al. studied psychiatrists' decisions to use psychoactive drugs in 40 "paper" cases[12] and found little agreement among physicians in their decisions, or in how they weighted the eight clinical variables in reaching a decision. Fisch et al. also investigated the prescribing of psychoactive drugs.[10] In a study of American and Swiss psychiatrists, he found that the level of agreement among the physicians generally was not above chance.

Rothert used a linear model to assess how physicians and patients make decisions about compliance with antihypertensive regimens.[44] Later, she and her colleagues modeled the weighting used by physicians in deciding whether to refer obese patients for further medical workup.[45,46] In the latter study, the rate of referral differed by medical specialty, and the most important factor was non-medical: the patient's desire for further consultation.

Richardson and colleagues also investigated physician referral policies. They analyzed 211 obstetricians' decisions about whether to refer high-risk obstetrical patients based on "paper" cases with ten medical and social risk factors.[42] They found considerable variation among physicians. Although medical factors were weighted the most heavily, other factors such as distance to the referral center and socioeconomic status were important for some.

Using a computer feedback model, Smith and I studied how physicians use factors such as patient wishes, family wishes, and prognosis in deciding whether to begin tube feeding in seriously ill patients.[51] Finding a great diversity of strategies, we used cluster analysis to define three groups of physicians with similar strategies.

Rovner and colleagues, studying how physicians decide which test to order in working up hypertension, also found considerable variation from one physician to the next.[47] Like Kirwan, they demonstrated that the judgments made in response to simulated cases closely resembled those made with actual patients.[48]

EDUCATIONAL EVALUATION

Linear models have proved useful, as in Thorndike's proposal, for determining what qualities evaluators consider important in judging trainees or applicants. Orkin and colleagues studied the weights faculty evaluators gave to different characteristics of anesthesiology residents,[36] and Clarke and I modeled how faculty judge applicants for surgical residency positions (see fig. 7). We later used the derived weighting scheme to screen subsequent applicants.[5] Recently, Young and colleagues determined the weights faculty members used to rate 441 applicants for an internal medicine residency program. They were able to explain a high percentage of the variation ($r^2 = 0.69$) on the basis of applicant characteristics.[64]

COGNITIVE FEEDBACK

In a further application of the lens model, Hammond showed that feedback of the cue weights, which he called "cognitive feedback," produced better results

than simple outcome feedback in learning complex judgmental tasks.[20] He developed a computer program that calculated students' weighting of cues from their decisions and then provided the students with a graphic display of their weights compared with the optimal weights. Students given cognitive feedback outperformed other students who received only outcome feedback (i.e., merely the correct answer to each problem). Moreover, other studies have suggested that when complex judgment tasks are needed, outcome feedback can even retard learning.[21]

To determine whether Hammond's findings regarding the superiority of cognitive feedback applied to medical diagnosis, I wrote a microcomputer program, titled FEEDBACK, to present simulated cases of urinary tract infection and to provide cognitive feedback. The program first analyzed students' responses over a series of cases and then displayed a comparison of weights calculated from their responses with the weights derived from a large series of clinical cases where the true diagnosis was known.[61] Our studies showed that cognitive feedback was highly effective for teaching diagnostic accuracy, and that it was superior to outcome feedback alone. We found also that the greater accuracy resulting from cognitive feedback was accompanied by the students' convergence on the correct weights and by a greater similarity of strategies among group members.

Recently, we adapted the program FEEDBACK to improve the diagnostic estimates of the student health physicians studied by Poses and colleagues.[38] Cognitive feedback improved their calibration on simulated cases, i.e., it dramatically improved the correspondence between predicted and actual frequencies of occurrence.[63] Moreover, cognitive feedback improved both diagnostic calibration and discrimination when these physicians diagnosed actual cases of patients seen in the student health clinic.[38]

OBSERVATIONS AND CONCLUSIONS FROM THE EXPERIMENTAL STUDIES

These medical applications have consistently found wide variations among physicians in how they use information to make diagnoses or treatment decisions. Differences in weighting explain much of this variation. In general, derived models of judgment have been more accurate in predicting later decisions than physicians' self-reported strategies.

Thus, a linear model calculated from actual clinical cases often differed from the physician's stated model and from the model derived from physicians' diagnostic estimates for "paper" cases. Further, the model calculated from clinical cases usually outperformed both the model calculated from the judge's decisions and the judge's stated model. Similarly, the model calculated from the judge's decisions usually outperformed his or her stated model when applied to the same cases. Simple models with unit weights and no interactions often do equally well.[7] Feedback of cue weights (cognitive feedback) improved teaching of complex diagnostic or decision strategies and surpassed outcome feedback in such situations.[20] Finally, the judgments made in response to "paper" cases have in several studies resembled those made with actual patients.[30,48]

Methods—Construction and Analysis of Simulated Cases

One can construct a linear model of a judge's decision either by observing real-life decisions or by recording decisions made about simulated cases. Calculating the model from actual cases may require a large number of cases and it may be hard to control for the large number of potential cues present. Simulated or "paper" cases allow control of the information presented, but provide no assurance that the judgments will be the same as in actual practice. As discussed earlier, several studies have found high correlations between judgments regarding "paper" cases and real-life judgments,[30,48] but this has not been tested over a wide range of situations.

The basic tasks in constructing simulated cases involve choosing an appropriate outcome, selecting the salient variables, deciding how to present them, choosing the range and levels of values for each variable, planning the underlying design for setting the value of each variable in each case, and selecting the method of analysis. Each of these steps can have a major influence on the results and the conclusions.

DESIGN OF THE CASES

Three approaches have been taken in creating simulated medical cases. Studies growing out of the tradition of the lens model[18] have been concerned with presenting realistic cases, representative of those ordinarily encountered ("representative" design). They often use continuous variables and may select realistic cases from randomly generated combinations of cue values. Random variation of cue levels is used in Hammond's computer program POLICY[22] and the later microcomputer version POLICY-PC by Rohrbaugh.[43] Advantages of this approach are realism of cases and flexibility of design. Disadvantages are the large number of cases required, increasing with the number of variables. The design may not be balanced, and interactions among the variables may not be measurable and may interfere with estimation of main effects.

Interactions occur when the weighting of one cue depends on the status of another. For example, a physician might ignore white blood cells in the urine when epithelial cells are present but consider them of great importance when epithelial cells are absent. It is not

clear whether it is important to look for and measure such interactions in most diagnostic settings.

A second approach is to use a full factorial design where all possible combinations of all variables are represented.[45,46] Advantages of this approach are that investigators can measure all interactions among variables and calculate weights precisely. Disadvantages include the large number of cases needed, the potential for creating unrealistic cases, and the inability to portray intercorrelated cues. For example, a full design with four variables at each of three levels requires 81 cases ($3 \times 3 \times 3 \times 3$). Such designs are most useful where the important cues are few and when precise measurement of weights and interactions is considered important.

A third approach is to employ fractional factorial designs.[1,14,37] These designs use a fraction of the cases in the full design but are devised so that each level of each variable is combined an equal number of times with each level of all other variables; however, not all the possible combinations are represented. These designs came into medical applications largely through the use of conjoint analysis, a nonparametric linear modeling technique widely used in marketing.[15,36,42,62]

Several early medical applications used fractional factorial designs.[41,42,47,61,62,63] The advantages of these designs are that they permit a smaller number of cases with more variables per case. Further, increasing the number of variables may not require more cases, calculating weighting is simpler, and, in some types of designs, first-order interactions can be estimated.[37] The disadvantages are that there is an even greater need to have uncorrelated cues, the design may generate unrealistic cases, and the smaller number of cases may lead to less accurate estimates of weighting. Fractional factorial designs are most useful when one wishes to portray a large number of generally unrelated variables. For examples of designs, see the articles by Plackett and Burman[37] and Addelman.[1]

SELECTION OF VARIABLES

The task in choosing variables is to cull the most important cues from those available to the decision maker. One may select cues by several methods. One may simply choose those cues that are of greatest interest. For medical decisions, one may review the literature to find the cues accepted as essential to diagnosis or treatment. One may employ consensus techniques such as nominative group process or the Delphi method to avoid missing important variables. The cues that people say are important to them, however, may not be those they actually use in making judgments, and textbooks may not yield the optimal

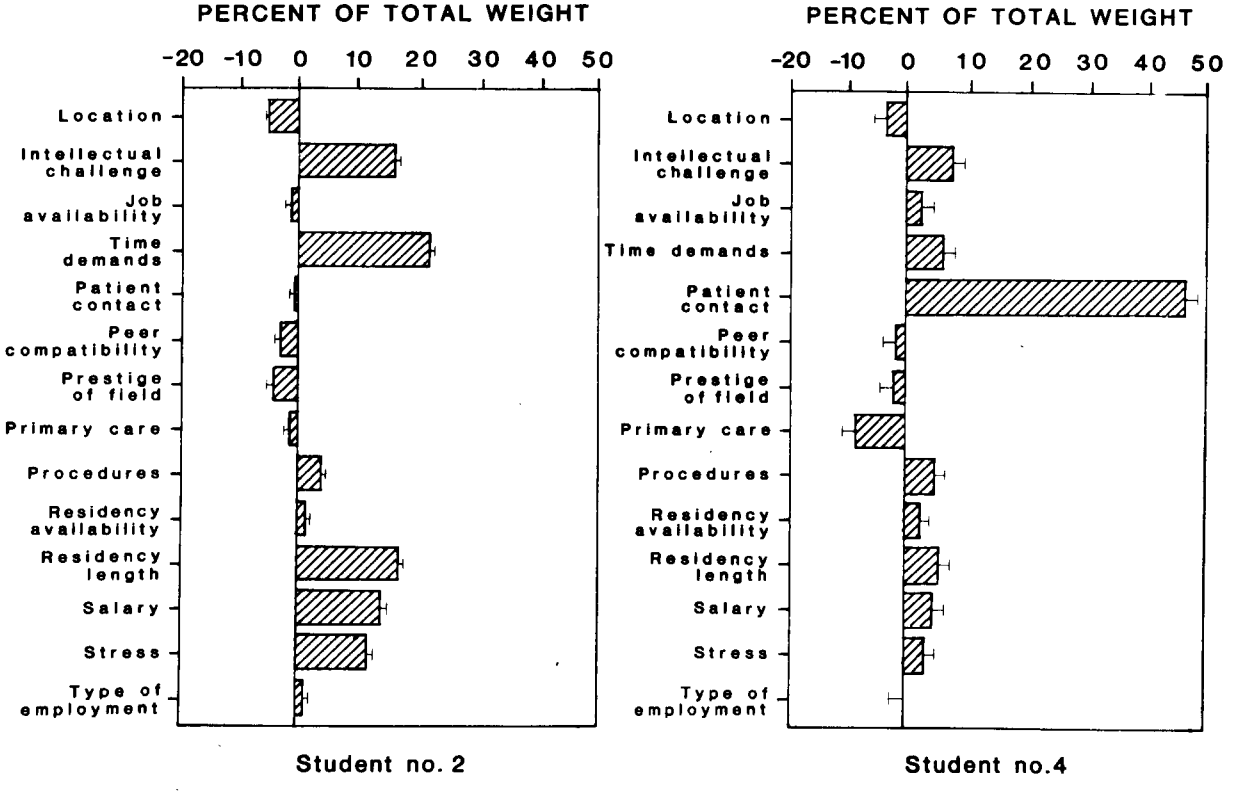

FIGURE 2. Differences in weighting of medical specialty characteristics by two students. The factors important to student 2 are of lesser importance for student 4, who gives the most weight to the presence or absence of patient contact. Averaging of weighting may disguise high variation among decision makers. Bars represent one standard error of the estimate.

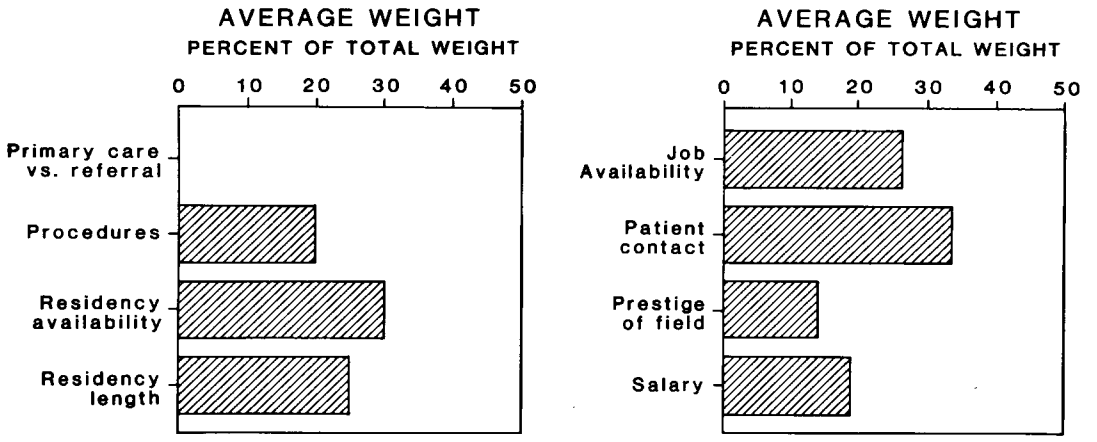

FIGURE 3. Illustration of the difficulty of knowing whether important or unimportant variables have been selected for paper cases. The figure shows the average weights calculated for six students in rating medical specialties. The first set of cases (*left*) used the four *least* important attributes ($r^2 = 0.89$) and the second set of cases used the four *most* important attributes ($r^2 = 0.94$). There was no indication from the results that the first study used relatively unimportant cues and that the second used the most important cues.

multivariate clinical predictors. One solution is to conduct pilot studies using simulated cases or paired comparisons to help reduce the number of cues.[16] Another approach is to select the cues by analyzing large numbers of actual cases to determine the optimal clinical predictors,[61] although this strategy may cause the modeler to omit variables that are considered important by the physicians but are weak predictors in the actual cases.

EFFECT OF CHANGES IN CUE SELECTION AND REPRESENTATION—AN ILLUSTRATION

Because little is known about the effect of differences in how cues are selected and presented, I asked six medical students to rate the desirability of several sets of descriptions of medical specialties to determine the effect of changes in cue selection and presentation on the students' weighting of many medical specialty attributes. After calculating the weight each student initially gave each variable, I then presented cases with different numbers of variables, with important variables missing, and with variation in the range of values for the variable. The results of these trials are discussed in the following four sections.

REASONS IMPORTANT VARIABLES MAY BE OMITTED. When a few variables must be selected from many to create simulated cases, there is always concern that one or more of the variables used by a judge may be omitted. These variables may have been overlooked, may have been considered but ruled unimportant, or may have been hard to portray in a "paper" case (e.g., "the patient appears toxic," "the patient seemed depressed").

A related problem is when one or more judges employ cues entirely different from those used by the majority. Here, consensus methods for selecting the best cues may exclude cues highly important to some of the subjects. For example, in considering medical specialty attributes, I asked the students to rate each specialty based on 14 characteristics such as salary level, job availability, amount of stress, peer esteem, and intellectual challenge. It was not difficult to find pairs of students where one gave the greatest weight to cues that were not at all important for the other (fig. 2). In attempting to reduce this array to three or four cues, one could falsely conclude that student's strategy employed particular cues, when, in truth, he would not use them at all when given a larger selection.

DIFFICULTY IN RECOGNIZING WHEN IMPORTANT VARIABLES ARE MISSING. Can we determine when important cues are left out of the simulation? I examined this question with the same six medical students by creating two different sets of hypothetical specialty descriptions. One set described the 16 specialties using the four variables that had received the *most* weight from these students (job availability, patient contact, salary, and prestige). The second set of 16 descriptions used the four variables that had received the *least* weight (primary care vs. referral practice, performance of procedures, availability of residencies, and length of residencies). The results revealed no clear marker that the one set of four variables had been less important that the other (fig. 3). The ranges and means of the ratings of the hypothetical specialties were similar, as was the variance explained ($r^2 = 0.94$ for the four most important and $r^2 = 0.89$ for the four least important). This example suggests we cannot rely on the way judges employ cues in simulated cases to indicate whether the cues presented were those they would weight most heavily given a larger selection. Furthermore, there is good evidence that irrelevant cues may affect decisions,[11] so a simulation including only relevant variables may not be an adequate model of the decision task.

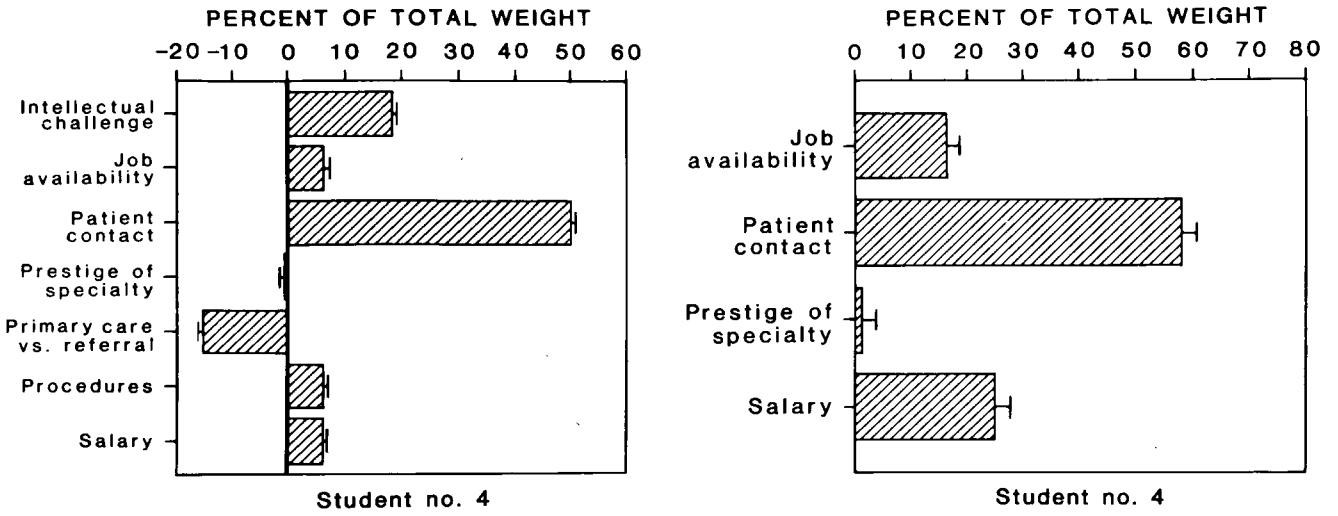

FIGURE 4. Effect of changing the number of variables. Weights of the same four variables for student 4 calculated from cases using different numbers of two-level variables: seven, four and 14 (in fig. 2, *right*). The order of importance of the variables is preserved but both relative and absolute weights differ. Bars represent the standard error of the estimate.

One must take great care in choosing variables because it may be difficult or impossible to tell whether crucial variables have been left out of the case simulations. In clinical medical studies the process of selecting variables should be reproducible so that it can be confirmed by other investigators. Reliable procedures for assuring that important cues have not been excluded need to be developed.

EFFECTS OF CHANGES IN THE NUMBER OF VARIABLES. Another problem is that the number of variables used can affect their weighting. Student 4, whose weighting of specialty characteristics is shown on the right side of figure 2, was given two new sets of simulated specialties that differed in the numbers of variables presented. One set had seven of the original 14 variables at two levels in 16 cases; another had four of these seven variables at two levels (fig. 4). Thus, weights for the same four variables were calculated under three different conditions; first, as four of 14 variables; second, as four of seven variables; and third, as the only four variables presented. The results showed that although the relative orders of weighting of the four variables were similar in these three sets, the relative and absolute weights for each of the four variables differed considerably. Salary, for example, represented an important factor with over 20% of total weight in the four-variable example (fig. 4), while in the 14-variable example it was of lesser importance and accounted for only 6% of total weight.

REPRESENTATION OF VARIABLES

There is a great deal of latitude in choosing how to describe and present cues within each case descrip-

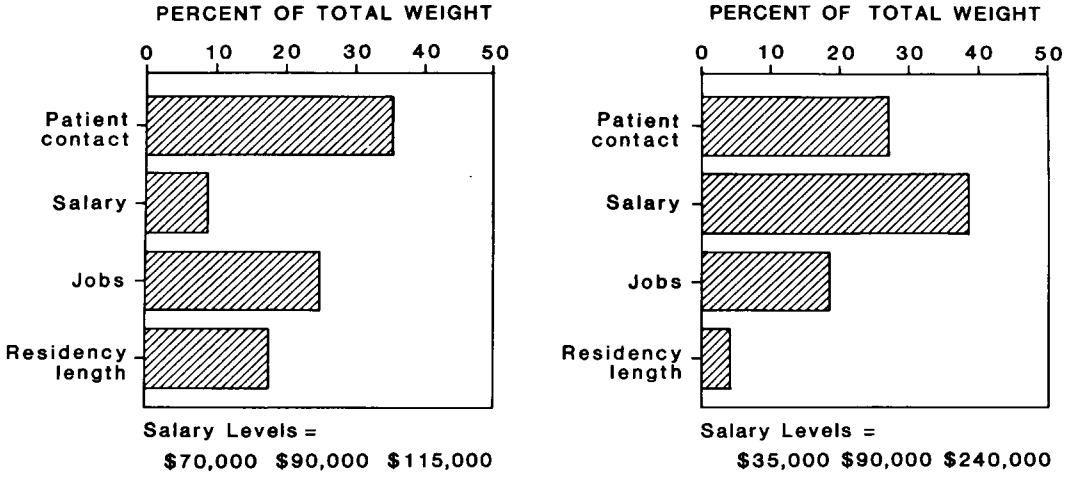

FIGURE 5. Effect on weighting of changing the range of a variable. In rating medical specialties, students gave the least weight to salary when the three levels were $70,000, $90,000 and $115,000. When the range was broadened to $35,000, $90,000, and $240,000, salary became the most heavily weighted both on the average and for each of the six students.

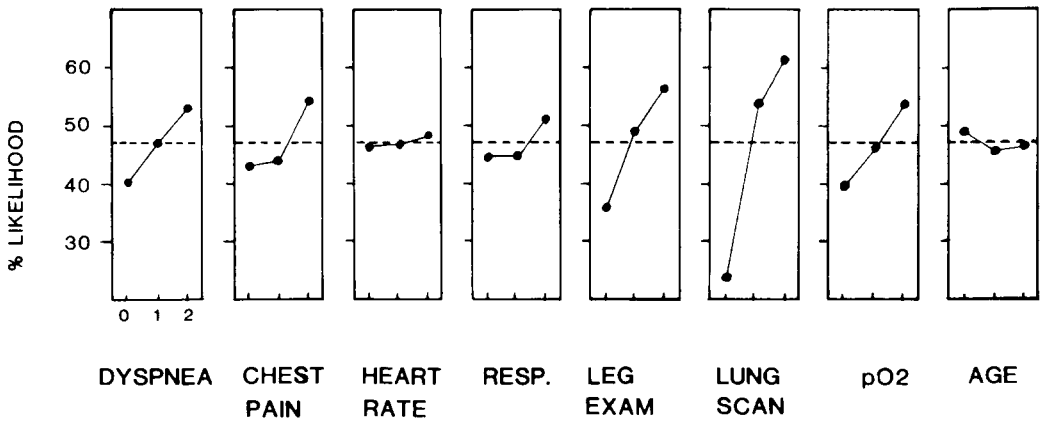

FIGURE 6. Utilities of clinical variables as determined by conjoint measurement. Mean likelihood of pulmonary embolus shown for each of three levels for eight cues. This type of display shows not only the weight given the cue (the difference between the highest and lowest value is equal to the β coefficient in regression) but also the linearity of the response.

tion, and these choices may affect the results. For example, the range of values over which the cue is varied can affect weighting. In the study of students' specialty choices, salary received little weight from students. Figure 5 shows what happened when I varied the range of possible salaries in two different sets of 18 cases with all variables at three levels. In the first sets of 18 cases with all variables at three levels. In the first set of 18 cases, salary was set at either $70,000, $90,000, or $115,000, representing the low, middle, and high levels. Students gave salary the lowest weight of all four variables in this set of cases. When the salary levels were set at $35,000, $90,000, and $240,000, salary became the most important variable for all six students, whether analyzed individually or as a group. Thus, the range over which a cue varies can affect the conclusions reached about its importance.

It is important to have reproducible techniques for determining what levels or ranges to use. There is no problem when clearly defined levels occur in the real-life decision setting, such as when the cue is naturally dichotomous (as with many laboratory tests) or when standard cut-off values are widely used and accepted. If the natural distribution of the variable is known, levels can be set using percentiles or other measures of distribution. Levels can also be determined empirically through a study of actual cases. Finally, one can experimentally measure the sensitivity of the conclusions to changes in variable levels. If it is the purpose of a study to measure the relative importances of several cues, then attention to the range and distribution of cue values is critical.

An advantage of using factorial designs with three or more levels is that one can use the mean values for each of the cue levels to determine whether response is linear over the range. It may be useful to study some variables in detail with multiple levels to detect whether there are important thresholds or other nonlinear aspects to the way the variable is used. Figure 6 shows the mean likelihood of each cue level as used by physicians in diagnosing simulated cases of possible pulmonary embolus.[62] It is clear, for example, that the intermediate level for the lung scan variable (selected to represent a true 50% likelihood) was interpreted as

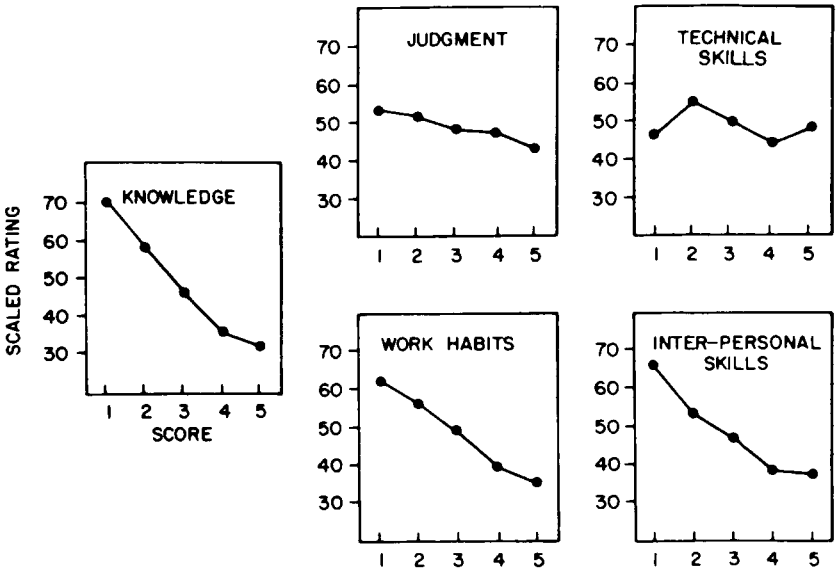

FIGURE 7. Mean rating of candidates for surgery residency programs according to the level of evaluation received for each of five dimensions. The five-level design permits examination of the linearity of response over a variety of values.

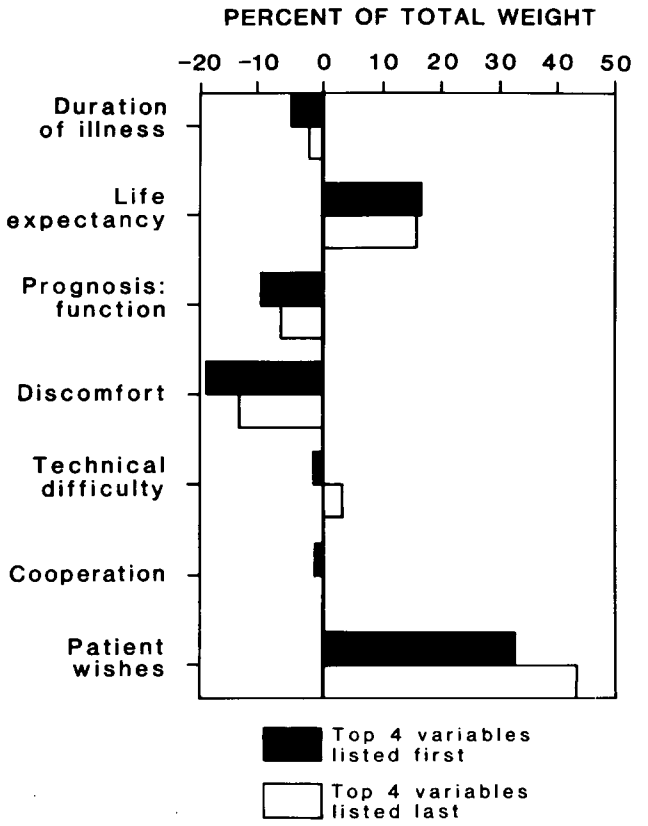

FIGURE 8. Effect of variable order on weighting. When the first four variables shown were presented in the first paragraph of each written vignette (39 respondents, 27 cases each), they received greater weight than when presented in the second paragraph (46 respondents). The same was true for the bottom three variables.

being much closer to the positive end. Designs with more levels can provide even more detailed information, as in the study of factors influencing the selection of surgical residents (fig. 7), but require fractional factorial designs because of the large number of cases that would be generated in a full design.

Another feature of case design that can affect weighting is the order in which variables are presented. This effect becomes more pronounced as the vignette gets long or complicated. Smith and I examined the effect on weighting of changing the order in which variables are presented. In a study of how physicians decide whether to begin tube feeding,[51] we presented seven variables in cases two paragraphs in length. We randomized the 85 participants into two groups. One group responded to cases with one paragraph printed first (39 respondents); the other responded to cases with the order reversed (46 respondents). As shown in figure 8, the same variables received greater weight when presented in the paragraph that appeared first in the written case (DG Smith, RS Wigton, unpublished data).

Other influences on variable weighting are the length of the description of the variable, the wording, and the format of the case itself. Also, subjects may not always interpret variables the way the experimenter intends. Ravitch and associates examined physicians' interpretations of cues describing cancer risk, symptom severity, and osteoporosis risk. The authors found that the cues were not interpreted in the same way by all participants.[41]

METHOD OF ANALYSIS

The design of the study and the outcome measure used largely determine the method of analysis. Multiple linear regression is used where the outcome is a continuous variable with a linear relationship to the case variables. Analysis of variance and dummy variable regression provide the most information when the variables are expressed in discrete levels, as in a factorial design.[15,42,60] Logistic regression is preferred for dichotomous outcomes.[3,60] Cox regression might be the best technique if the outcome is the time to an event (although I am not aware of any policy capturing studies that have used it). Logit transformation of the outcome before performing regression analysis may be required if the outcome is a probability estimate.[63] Conjoint analysis (analysis of variance incorporating a monotone transformation based on a goodness-of-fit measure) can be used to determine weighting when the cases are to be ranked.[15,36,42]

INDIVIDUAL VS. GROUP LEVEL OF ANALYSIS

Should strategies be analyzed at the level of the individual or at the level of the group? Some feel that aggregate or group strategies are meaningless, regardless of how they are calculated[23] because they may combine strategies that are inconsistent with one another (i.e., a strong positive weight for a factor averaged with a strong negative weight would lead to the conclusion that the factor was not important).

On the other hand, group weighting appears to be reproducible and predictive and has been used extensively in marketing research[15] and in several medical studies.[45,46,64] Two methods that have been used to describe group strategies are analyzing the pooled data of all judges, and analyzing each judge's data separately and then averaging the weights. Alternatively, one can analyze each judge's data separately and characterize the distribution of strategies either in terms of simple descriptive statistics (median, percentiles) or through more sophisticated methods such as cluster analysis.[32,41,50] In view of the marked heterogeneity of physician strategies found in many of the studies, this is an important methodologic consideration.

Problems and Challenges in the Application of Linear Models

Despite many innovative designs and applications, there remain major limitations to how well judgment

analysis can model real-life decisions. The constraints imposed by the analytic methods make the simulated cases unlike real-life decision settings in several ways. The simulated cases contain only a fraction of the variables present in the real-life decision setting. The information available in real-life decisions usually includes many redundant and intercorrelated cues, a feature not often included in the simulated cases. Design constraints, particularly with factorial designs, may cause variables to be presented in the simulated cases as dichotomous, while in the actual decision setting they are continuous. Finally, acquisition of information in the medical setting often is sequential and incomplete, whereas simulations present all cue values simultaneously whether or not they are requested.

Future Directions

There is a growing body of research in medicine using clinical prediction rules derived from multivariate analysis of patient data. A recent review described 16 such rules derived by multivariate analysis that were found in review of four clinical journals from 1981 through 1984.[60] These studies are providing much-needed data on the actual relationship of cues to the clinical diagnoses and decisions. It is not clear what is the best way to apply these multivariate rules to clinical medicine. Should they be taught to physicians,[24] incorporated into computer prediction aids,[40] or incorporated into the test ordering procedures? Studies of how physicians best learn and utilize linear rules will be of great importance in applying the results of these studies. Initial results with cognitive feedback are very encouraging.[33,61]

Another area in medical education is the development of simulations of diagnostic and therapeutic decisions. Such simulations have potential not only in teaching appropriate weighting but also in increasing the accuracy of probability estimates.[39,63]

Several limitations of current designs will need to be addressed to improve the usefulness and realism of simulations. Cluster analysis and clustering of variables in simulations may allow models to portray multiple cues with redundant information, a feature common in medical decisions. Methods for increasing the number of variables in the simulation or for interactively reducing the number from an initially large field may help avoid missing important variables. Microcomputer programs will be important in incorporating these features in simulations.

There are reasons to believe linear models will be the most productive route for studying medical decisions. They consistently match or outperform experts in prediction in their own areas. The insights gained from this type of analysis, unlike those gained from other models, have been shown to enhance learning of judgment tasks. The basic design of the task is very similar to one of the most common judgment tasks encountered in medical practice: a decision or prediction based on a combination of many known factors where the physicians must select and weight the factors to arrive at a judgment.

Linear models will be valuable in studying why there is such variation in physician judgments and in teaching physicians to be better diagnosticians and prognosticators. The growing use of multivariate models in medicine and the ready availability of computers should further increase the usefulness of linear models in studying and aiding medical diagnosis.

The author is grateful to Drs. Randall Cebul, Roy Poses, Robert Centor, Kenneth Hammond, Thomas Tape, Marilyn Rothert, David Rovner, and David Smith for their helpful comments and suggestions, to JoAnna Nicolas for assistance with data analysis, to Kashinath Patil, PhD, for statistical advice, and to Vicki Hamm for preparing the manuscript.

References

1. Addelman S: Orthogonal main-effect plans for asymmetrical factorial experiments. Technometrics 4:21–46, 1962
2. Anderson G, Bombardier C: Estimating disease activity in rheumatoid arthritis. Med Decis Making 4:469–487, 1984
3. Centor RM, Witherspoon JM, Dalton HP, et al: The diagnosis of strep throat in adults in the emergency room. Med Decis Making 1:239–246, 1981
4. Cebul RD, Poses RM: The comparative cost-effectiveness of statistical decision rules and experienced physicians in pharyngitis management. JAMA 256:3353–3357, 1986
5. Clarke JR, Wigton RS: Development of an objective rating system for residency applications. Surgery 96:302–306, 1984
6. Dawes RM: A case study of graduate admissions: application of three principles of human decision making. Am Psychologist 26:180–188, 1971
7. Dawes RM, Corrigan B: Linear models in decision making. Psychol Bull 81:95–106, 1974
8. Einhorn HJ: Expert measurement and mechanical combination. Organ Behav Hum Perform 7:86–106, 1972
9. Fisch H-U, Hammond KR, Joyce CRB, et al: An experimental study of the clinical judgment of general physicians in evaluating and prescribing for depression. Br J Psychiat 138:100–109, 1981
10. Fisch H-U, Gillis JS, Daguet R: A cross-national study of drug treatment decisions in psychiatry. Med Decis Making 2:167–177, 1982
11. Gaeth GJ, Shanteau J. Reducing the influence of irrelevant information on experienced decision makers. Organ Behav Hum Perform 33:263–282, 1984
12. Gillis JS, Lipkin JO, Moran TJ: Drug therapy decisions, a social judgment analysis. J Nerv Ment Dis 169:439–447, 1981
13. Goldberg LR: Man versus model of man: a rationale, plus some evidence for a model of improving on clinical inferences. Psychol Bull 57:116–131, 1960
14. Green PE, Carroll JD, Carmone FJ: Some new types of fractional factorial designs for marketing experiments. Res Marketing 1:99–122, 1978
15. Green PE, Rao VR: Conjoint measurement for quantifying judgmental data. J Marketing Res 8:355–363, 1971
16. Green PE, Carroll JD, Goldberg SM: A general approach to product design optimization via conjoint analysis. J Marketing 45:17–37, 1981

17. Hammond KR: Probabilistic functioning and the clinical method. Psychol Rev 62:255–262, 1955
18. Hammond KR, Hursch CJ, Todd FJ. Analyzing the components of clinical inference. Psychol Rev 71:438–456, 1964
19. Hammond KR: The Psychology of Egon Brunswik. New York, Holt, Rinehart and Winston, 1966
20. Hammond KR: Computer graphics as an aid to learning. Science 172:903–908, 1971
21. Hammond KR, Summers DA, Deane DH: Negative effects of outcome feedback in multiple cue probability learning. Organ Behav Hum Perform 9:30–34, 1973
22. Hammond KR, Stewart TR, Brehmer B, et al: Social judgment theory. *In:* Kaplan MF, Schwart S (eds): Human Judgment and Decision Processes. New York, Academic Press, 1975, pp 271–312
23. Hammond KR, McClelland GH, Mumpower J: Human Judgment and Decision Making. New York, Praeger, 1980
24. Hickam DH, Sox HC: Teaching medical students to estimate probability of coronary artery disease. J Gen Intern Med 2:73–77, 1987
25. Hoffman PJ: The paramorphic representation of clinical judgment. Psychol Bull 57:116, 1960
26. Holzman GB, Ravitch MM, Metheny W, et al: Physicians' judgments about estrogen replacement therapy for menopausal women. Obstet Gynecol 63:303–311, 1984
27. Hursch, C, Hammond KR, Hursch JL: Some methodologic considerations in multiple cue probability studies. Psychol Rev 71:42–60,1964
28. Joyce CRB, Last JM, Weatherall M: Personal factors as a cause of difference in prescribing by general practitioners. Br J Prev Soc Med 21:170–177, 1967
29. Joyce CRB, Hammond KR: Judgment analysis in clinical trials. How to save the baby from the bathwater. Controlled Clin Trials 5:307–308, 1984
30. Kirwan JR, Chaput de Saintonge DM, Joyce CRB, et al: Clinical judgment in rheumatoid arthritis. I. Rheumatologists' opinions and the development of "paper patients." Ann Rheum Dis 42:648–651, 1983
31. Kirwan JR, Chaput de Saintonge DM, Joyce CRB, et al: Clinical judgment in rheumatoid arthritis. II: Judging current disease activity in clinical practice. Ann Rheum Dis 42:648–651, 1983
32. Kirwan JR, Chaput de Saintonge DM, Joyce CRB, et al: Clinical judgment in rheumatoid arthritis. III. British rheumatologists' judgments of change in response to therapy. Ann Rheum Dis 43:686–694, 1984
33. Kirwan JR, Chaput de Saintonge DM, Joyce CRB, et al: Inability of rheumatologists to describe their true policies for assessing rheumatoid arthritis. Ann Rheum Dis 45:156–161, 1986
34. Kruskal JB: Analysis of factorial experiments by estimating monotone transformations of the data. J Statistical Soc Series B 27:251–263, 1965
35. Luce RD, Tukey JW: Simultaneous conjoint measurement: a new type of fundamental measurement. J Math Psychol 4:1–27, 1964
36. Orkin FK, Greenhow DE: A study of decision making. Anesthesiology 48:267–271, 1978
37. Plackett RL, Burman JP: The design of optimum multifactorial experiments. Biometrika 33:305–325, 1946
38. Poses RM, Cebul RD, Wigton RS, et al: Feedback on simulated cases to improve clinical judgment (abstr). Med Decis Making 6:274, 1986
39. Poses RM, Cebul RD, Collins M, et al: The accuracy of experienced physicians' probability estimates for patients with sore throats. JAMA 254:925–929, 1985
40. Pozen MW, D'Agostino RB, Selker HP, et al: A predictive instrument to improve coronary-care unit admission practices in acute ischemic heart disease: a prospective multicenter clinical trial. N Engl J Med 310:1274–1278, 1984
41. Ravitch MM, Methany WP, Holzman GB, et al: Effects of physicians' interpretations of clinical cues and self-reported weights on modeling clinical judgment (abstr). Med Decis Making 3:368, 1983
42. Richardson DK, Gabbe SG, Wind Y: Decision analysis of high-risk patient referral. Obstet Gynecol 63:496–501, 1984
43. Rohrbaugh JR: Policy PC: Software for Judgment Analysis. New York, Executive Decision Services, 1986
44. Rothert ML: Physicians' and patients' judgments of compliance with a hypertensive regimen. Med Decis Making 2:179–195, 1982
45. Rothert ML, Rovner MD, Elstein AS, et al: Differences in medical referral decisions for obesity among family practitioners, general internists, and gynecologists. Med Care 22:42–53, 1984
46. Rovner DR, Rothert ML, Holmes MM, et al: Rationale for physicians' decisions to refer obese patients. Med Decis Making 5:279–292, 1985
47. Rovner DR, Rothert ML, Holmes MM: Validity of structured cases to study clinical decision making (abstr). Clin Res 34:834a, 1986
48. Rovner DR, Rothert ML, Holmes MM, et al: Validating case vignettes with clinical practice: the case of urinary tract infection (abstr). Med Decis Making, 6:272, 1986
49. Slovic P, Rorer LG, Hoffman PJ: Analyzing the use of diagnostic signs. Invest Radiol 6:18–26, 1971
50. Slovic P, Lichtenstein S: Comparison of Bayesian and regression approaches to the study of information processing in judgment. Organ Behav Hum Perform 6:649–744, 1971
51. Smith DG, Wigton RS: Modeling decisions to use tube feeding in seriously ill patients. Arch Intern Med 147:1242–1245, 1987
52. Smith DG, Wigton RS: Use of conjoint analysis to determine how physicians weight ethical considerations in making clinical judgments (abstr). Med Decis Making 3:376, 1983
53. Stewart TR, Joyce CRB, Lindell MK: New analyses: application of judgment theory to physicians' judgments of drug effects. *In:* Hammond KR, Joyce CRB (eds): Psychoactive Drugs and Social Judgment. New York, Wiley Interscience, pp 249–262
54. Stewart TR, Joyce CRB: Increasing the power of clinical trials through judgment analysis. Med Decis Making 8:33–38, 1988
55. Thorndike EL: Fundamental theorems in judging men. J Appl Psychol 2:67–76, 1918
56. Tucker LR: A suggested alternative formulation in the development of Hursch, Hammond and Hursch and by Hammond, Hursch, and Todd. Psychol Rev 71:528–530, 1964
57. Ullman DG, Doherty ME: Two determinants of the diagnosis of hyperactivity: the child and the clinician. Adv Devel Behav Pediatr 5:167–217, 1984
58. Von Winterfeldt D, Edwards W: Decision Analysis and Behavioral Research. Cambridge, Cambridge University Press, 1986
59. Wallace HA: What is in the corn judge's mind? J Am Soc Agronomy 15:300–304, 1923
60. Wasson JH, Sox HC, Neff RK, et al: Clinical prediction rules: applications and methodological standards. N Engl J Med 313:793–799, 1985
61. Wigton RS, Patil KD, Hoellerich VL: The effect of feedback in learning clinical diagnosis. J Med Educ 61:816–822, 1986
62. Wigton RS, Hoellerich VL, Patil KD: How physicians use clinical information in diagnosing pulmonary embolism: an application of conjoint analysis. Med Decis Making 6:2–11, 1986
63. Wigton RS, Poses RM, Cebul RD. Teaching a linear decision rule for the diagnosis of streptococcal pharyngitis using computer feedback. Presented at the Seventh Annual Meeting of the Society for Medical Decision Making, October 1985
64. Young MJ, Woolliscroft JO, Holloway JJ. Determining the policies of a residency selection committee. J Med Educ 61:835–837, 1986