A Decision Analytic Method for Scoring Performance on Computer-Based Patient Simulations

Stephen M Downs[*], Charles P Friedman[†], Farah Marasigan[*], Gary Gartner[*]
[*]University of North Carolina at Chapel Hill
[†]University of Pittsburgh

As computer based clinical case simulations become increasingly popular for training and evaluating clinicians, approaches are needed to evaluate a trainee's or examinee's solution of the simulated cases. We developed a decision analytic approach to scoring performance on computerized patient case simulations.

We developed decision models for computerized patient case simulations in four specific domains in the field of infectious disease. The decision models were represented as influence diagrams. A single decision node represents the possible diagnoses the user may make. One chance node represents a probability distribution over the set of competing diagnoses in the simulations. The value node contains the utilities associated with all possible combinations of diagnosis and disease. All relevant data that the user may request from the simulation are represented as chance nodes with arcs to or from the diagnosis node and/or each other. Probabilities in the decision model were derived from the literature, where available, or expert opinion. Utilities were assessed by standard gamble from clinical experts.

The process of solving computer based patient simulations involves repeated cycles of requesting data (history, physical examination or laboratory) and receiving these data from the simulations. Each time the user requests clinical data from the simulation, the influence diagram is evaluated with and without an arc from the corresponding chance node to the decision node. The difference in expected utility between the two solutions of the influence diagram represents the expected value of information (VOI) from the requested clinical datum. The ratio of the expected VOI from the data requested and the expected value of perfect information about the diagnosis is a normative measure of the quality of each of the user's data requests.

This approach provides a continuous measure of the quality of the user's data requests in a way that is sensitive to the previous data collected. The score distinguishes serious from minor misdiagnoses. And the same influence diagram can be used to evaluate performance on multiple simulations in the same clinical domain.

INTRODUCTION

Interest in the use of computer based patient simulations in the training and evaluation of clinicians has grown in recent years because computer simulations offer interactivity and evolution of clinical problems over time in a way that is impossible with a paper based test. However, the field is only now developing rigorous methods for measuring a clinician's solution of a computer based case. [1, 2]

If clinical simulations are going to realize their potential for performance assessment, clinicians completing a lengthy simulated case must receive more than a single binary score based on whether the diagnosis was correct. A more sensitive scoring approach would measure the quality of the clinician's judgment throughout the evolution of the case.

This is difficult because evaluation of a clinician's judgment is context sensitive. Even in the same simulated case, an invasive test may be justified when certain information has been discovered earlier in the case, but inappropriate before that information is known. Approaches that do not take context into account also tend to reward novices for being thorough while penalizing experts for their efficiency.[3]

Scoring simulated cases as right or wrong based on the diagnosis also neglects that some misdiagnoses are worse than others. For example, when a treatable infection or an untreatable malignancy are both possibilities, it may make sense to treat the possibility of an infection even when malignancy is more likely.

Meaningful performance assessments should take into account the information previously gathered from the case, the relative seriousness of possible misdiagnoses, and the potential ability of each information request to improve the diagnosis. We have developed a system that utilizes the principles of decision analysis to measure these facets of a clinician's reasoning throughout a simulated case. This paper describes the use of decision models, represented as influence diagrams, to evaluate clinicians' solutions of simulated patient cases. We present the results of a pilot study which exemplifies the type of information this approach can yield.

METHODS

In the process of completing a patient simulation, the clinician is presented with a set of presenting

1091—8280/97/$5.00 © 1997 AMIA, Inc.

symptoms. The clinician evaluates the case by requesting information from the simulation, receiving this information, evaluating it, and requesting additional information. This cycle is repeated until a diagnosis is made. Each turn of the cycle is used as an opportunity to evaluate the clinicians information request.

Decision Analytic Approach

To score each information request, we utilize the decision analytic concept of expected value of information (VOI). Decision analysis is a method for comparing the relative merits of alternative actions based on the expected value of the possible outcomes of those actions.[4] Decision analytic models represent the alternative courses of action the decision maker may take, the probabilistic relationships between those actions and the possible resulting outcomes, and a quantitative representation of the relative desirability, or utility, of each outcome. The product of the probability and the utility of each outcome, summed over all the possible outcomes, is the basis for comparing alternative actions. The value of any course of action is measured by this expected utility.

Prior to choosing a course of action, a decision maker may collect additional information relevant to the decision. This information may alter the probabilities of certain outcomes and, thereby, change which course of action he will choose. For example, a blood test may suggest or exclude a particular diagnosis depending of whether the test is positive or negative. Any information that can potentially change which course of action is best has some potential value that can be measured by the change in the expected value of the decision. We call this change the expected VOI (Chap 5 in [4]), and it is the basis on which we score information requests from the clinician using the case simulation. To calculate the expected VOI, we have developed decision models using influence diagrams[5].

Developing the Influence Diagrams

An influence diagram is an directed, acyclic graph, i.e., a set of nodes incompletely connected by directed arcs.[5, 6] The nodes represent three types of variables in the decision model: chance nodes, decision nodes, and a value node. Each chance node represents a random variable and the probability distribution over its sample space. Arcs entering a chance node represent conditioning variables.

A decision node represents the set of alternative actions that may be taken at a given time. Arcs entering a decision node represent information that will be available at the time the decision is made. The value node represents the quantitative value, or utility, placed on the outcome of the decision. Arcs entering a value node come from those variables whose value affects the overall value or utility of the outcome.

We developed an influence diagram, for each of four clinical presentations (domains): fever and mental status changes, fever and rash, fever and cough, male urethral discharge. Each influence diagram contains between 21 and 27 chance nodes, one decision node, and one value node. The differential diagnosis for each included between 8 and 11 disease hypotheses. One chance node, called the disease node, represents a set of competing diagnostic hypotheses. Other chance nodes represent findings the clinician may request. Some "state" nodes represent states that may be present but are not directly observable. The probabilistic relationships between these nodes is shown by directed arcs between them. The differential diagnoses and relevant findings for each clinical presentation were obtained by review of the medical literature and expert opinion.

Figure 1 shows a simplified influence diagram representing the evaluation of a patient with fever and cough. The DISEASE node represents two competing possibilities, Bacterial Pneumonia or Viral Infection. Each is associated with a prior probability. Other chance nodes include the NEUTROPHILS node which may be Elevated or Normal. The probability that the neutrophils are elevated depends on whether the disease is bacterial or viral. Thus, there is a different probability of elevated neutrophils for each value of the DISEASE node. In epidemiological terms, this is the sensitivity of elevated neutrophils for bacterial disease. This relationship is depicted by the arc between DISEASE and NEUTROPHILS.

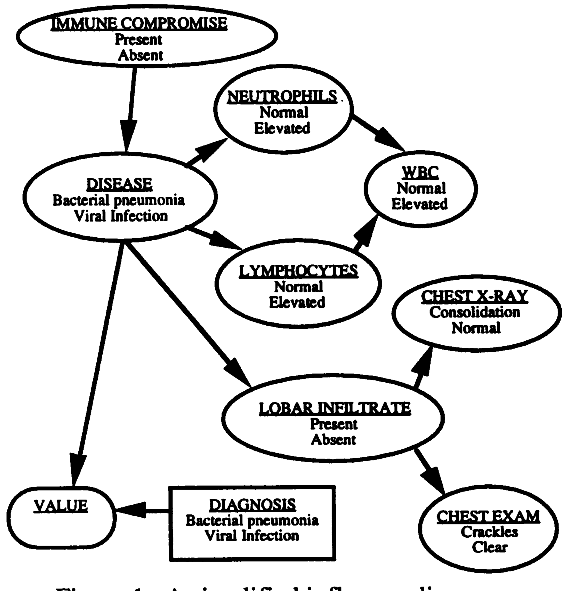

Figure 1. A simplified influence diagram.

Figure 1 also illustrates that some nodes representing findings may be derived from other nodes. In this case, the WBC, or white blood cell count, depends on both the neutrophil and the lymphocyte count. If the clinician requests a WBC without a differential count, the probability that the count will be elevated depends on the probabilities that the neutrophil and the lymphocyte counts are elevated.

Figure 1 also shows how the interdepencies between findings can be represented with hidden "state" variables. The patient may have a LOBAR INFILTRATE in the lung, and this probability depends on the underlying disease. However, the clinician cannot directly observe the presence of a lobar infiltrate. Instead, she can observe the presence of crackles on chest exam or consolidation on chest x-ray. These are indirect measures of lobar infiltrate. Because they both measure the same underlying process, albeit imperfectly, they provide partially redundant information. So the VOI from a chest x-ray may be substantially less after the presence of crackles has been detected on clinical exam.

In general, risk factors such as immune compromise, foreign travel, underlying diseases or other exposures were represented as chance nodes with arcs directed into the DISEASE node (figure 1).

Estimating Probabilities. Once the structure of the influence diagram was developed, the probabilistic relationships between variables were obtained by review of textbooks and literature identified by searching the Medline database. Data on prevalence of diseases and sensitivity and specificity of findings were used to estimate model parameters. Where these were not available, subjective estimates of internists, pediatricians and infectious disease specialists were used.

Assessing Utilities. The value placed on the outcome of the simulation depends on the diagnosis the clinician makes and the disease the simulated patient has. For every combination of diagnosis and disease, the value node stores a utility. By convention, the utility is a number between 0 and 1, where 0 is assigned to the worst outcome and 1 to the best. In our simplified example (figure 1), the highest utility (1) is associated with viral infection correctly diagnosed because this is the milder condition and the patient would receive no unnecessary treatment. The lowest utility is associated with bacterial pneumonia misdiagnosed as viral infection. In this case, a potentially fatal disease with an effective treatment would be left untreated. Intermediate values go to the correct diagnosis of bacterial pneumonia, which is still more serious than viral infection even with treatment, and viral infection misdiagnosed as bacterial pneumonia, leading to unnecessary treatment.

Utilities were assessed by interview of practicing internists, using the standard gamble method.[7]

Simulation Authoring and Delivery

A computer based patient simulation authoring and delivery program was developed using Hypercard. Eight cases, two from each of the four influence diagrams, were developed. The cases were authored by an infectious disease specialist and were based on clinical cases in his experience.

In each simulated case, clinicians are provided a case presentation, containing name, age, sex, and chief complaint of the patient. The clinician then has the opportunity to request information from a hierarchical menu of history, physical examination, and laboratory items. For each request, a log file records the item requested and the value returned by the simulation program to the clinician.

Solving the Influence Diagrams

A scoring program was developed which uses influence diagrams to score the clinician's interaction with the patient simulation by determining the expected VOI from each finding the clinician requests. The influence diagrams were evaluated using the algorithms described by Shachter.[6] The algorithms were implemented in C on a Macintosh Quadra 650. For each finding the clinician requests, the algorithm calculates the expected VOI as described below. The result provided by the simulation is used to instantiate the corresponding variables, updating the influence diagram.

Before the clinician requests any findings, the scoring program calculates the expected value for each diagnosis in the decision node of the influence diagram. The highest expected value among the diseases becomes the expected value of the simulation at that point. The program then reads the log files from the simulation program.

A finding in the simulation may correspond to none, one, or many of the nodes in the influence diagram. For each finding the clinician selects, an arc is introduced into the influence diagram going from the node(s) corresponding to the finding to the diagnosis decision node. The expected value of the influence diagram is then recalculated. The difference between the expected value of the influence diagram before and after the arc is introduced is the expected VOI for the finding.

Next the value of the finding given by the simulation is used to update the probability distribution over the diseases in the disease node, using Bayes' theorem, and the node(s) corresponding to the finding is eliminated.

Pilot Test

In a small demonstration study, a convenience sample of fourth year students at the University of North Carolina at Chapel Hill and the University of Pittsburgh was recruited to solve the computer based patient simulations. Trace files were obtained and analyzed by the scoring program. The results of nine students who completed a case of cough and fever were analyzed.

RESULTS

Pilot Test

Table 1 illustrates a small portion of the tabular output of the scoring program. Each row in the table corresponds to an information request by the clinician. The first column, *Simulation Item*, is the finding requested from the simulation. The second column, *Influence Diagram Node*, is the name of the node(s) corresponding to the finding. An entry of "NullNode" means the finding requested has no corresponding node in the influence diagram. The third column, *Expected VOI*, is the expected value of the finding *before* the result of the information request is known. The *Result* column shows the value of the finding returned by the simulation program, and the *Optimal Diagnosis*, is the diagnosis with the highest expected value *after* the value of the finding is known.

Only findings that, depending on their value, can potentially change the optimal diagnosis have a non-zero VOI. The expected VOI for any finding varies, depending on the items requested before it and the information returned by the simulation. For example, if a chest exam revealed findings consistent with pneumonia, the expected value of a chest x-ray may go down because the information is redundant. A finding that has no relevance to the clinical situation will have no corresponding node in the influence diagram and no VOI.

Simulation Item	Influence diagram node	Expected Value of Info.	Result	Optimal Diagnosis
Chest x-ray	CXR-finding	0.0038	Focal Densities	Chlamydia
CBC	WBC	0.032	Decreased	TB
Differen-tial Cell Count	Polys, Lymphs	0.000845	Normal, Decreased	TB
Hemo-globin	NullNode	0	nil	TB

Table 1: Tabular output of the scoring program

Among nine students completing one simulated case of fever and cough, the average number of findings requested per student was 64 (range 41-116). Of the findings requested, on average 5 findings (95% CI: 4, 6) had a non-zero VOI, i.e., had the potential to change the leading diagnosis.

One summary score for the solution of a simulation is the average VOI over all the findings requested. This statistic reflects efficiency because it increases in value as high information items are requested and decreases as low information items are requested. The average expected VOI for all items requested was per simulation in our pilot study was 0.8x10^{-3} (range 0 - 2.1x10^{-3}). Among findings with a non-zero VOI, the average expected VOI was 10^{-2} (range 3.4x10 $^{-3}$ - 3x10 $^{-2}$). The VOI is interval valued and has linear properties so the scale may be multiplied by any positive number or added to any constant without changing the relative values of the finding.[8, 9] So these values could be scaled up to make them more readily interpretable This also means that the values of information from one influence diagram can be scaled to be comparable to those from any other.

Figure 2 shows a graphical representation of the progress of one subject through the simulation. The horizontal axis shows the number of the information item requested during the simulation, and the vertical axis shows the item's expected VOI. This graph shows, at a glance, the points at which the data with positive VOI are obtained and their relationship to the type of data gathered (e.g., history, examination, or laboratory) or points of access to paper or computer-based information sources. In figure 2, for example, the laboratory data requested had the highest expected VOI.

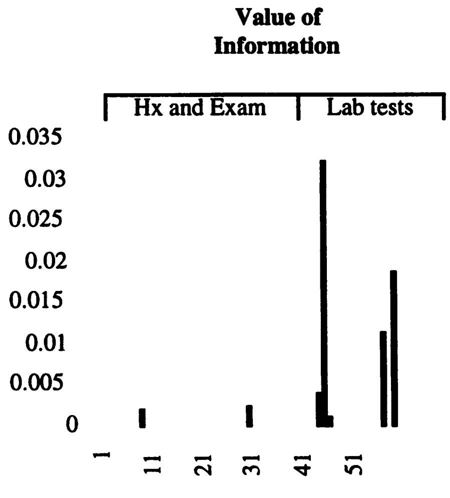

Figure 2: Graphical output of the scoring program.

DISCUSSION

We have developed a method for assessing performance on computer based clinical simulations using the decision analytic concept of expected VOI. The use of decision theory to assess performance offers several advantages. Because the theory uses probabilistic inference, relationships in the model can be derived from the scientific literature whenever such data exist. Fundamental epidemiological concepts such as prevalence, sensitivity, and specificity are explicitly represented in the influence diagrams. Moreover, the decision analytic approach is quite general. For example, it can be applied to simulations in which the clinician is asked to choose therapy.

The use of utilities can potentially take into account not only the sensitivity and specificity of a finding, but also the risks and costs of a test and the risks and costs of possible misdiagnoses. All of these concepts are weighed into the expected VOI measure.

This approach also evaluates every step in the solution of the simulated case, not just the correctness of the answer. In fact, when a case is very atypical but evaluated appropriately, the algorithm may produce a high score even if the answer is wrong. By scoring each step in the solution of a simulated case, the VOI provides a metric by which the diagnostic process can be dissected. It becomes possible to compare the information seeking behavior of experts and novices or to examine the effect of external information sources on data gathering, for example.

The approach can also generate summary scores. For example, the sum of VOI from all findings requested will reflect the thoroughness of the clinician, but the average VOI expresses the efficiency with which the case was done

The use of influence diagrams has some practical advantages as well. One influence diagram can be used to evaluate many simulations of different diseases. As long as the starting point of the simulations is essentially the same, any disease in the differential diagnosis can be represented. Moreover, using simulation techniques, new simulated cases can be derived automatically from the probabilistic relationships in the influence diagrams, a normally expensive and time consuming process.

Our pilot work suggests some challenges and unanswered questions. We have found that the process of generating the influence diagrams is difficult, and time consuming. Often critical data are not available in the literature, and subjective estimates are necessary. This may be offset somewhat by the relatively low marginal cost of creating new simulated cases once the influence diagram has been built.

Our utilities were obtained by standard interview techniques with medical experts, but utilities reflect individual preferences.[9] Decision analysis does not offer a method for obtaining "consensus" utilities. Measuring how sensitive the expected VOI is to changes in utilities is one objective of our ongoing work.

We found that relatively few (<10%) of the findings requested had a non-zero expected VOI. This may result from several characteristics of the specific diagnostic problem. For example, if the probability of a particular diagnosis is very low, many data supporting it must be obtained before it becomes the optimal hypothesis. Similarly, if one diagnostic possibility is easily and effectively treated, but quite dangerous if left untreated, it may remain the optimal diagnosis even when it is not the most likely diagnosis. The sensitivity of the expected VOI measure to the underlying probabilities and utilities more accurately reflects clinical reality. However, it suggests that selecting the right diagnostic problems will determine if enough findings have positive VOI.

A decision analytic approach to scoring clinical simulations appears to have potential applications in the study of clinical decision making processes as well a possible role in examinations of testing agency. While there are several advantages to a decision theoretic approach, many questions remain to be answered. We have demonstrated a proof of concept are conducting studies to explore the validity and reliability of the approach.

Acknoledgements

This study was supported by Grant R01-LM04843 from the National Library of Medicine.

REFERENCES

1.Friedman, C. and S. Downs, *Alternatives to current practice: Decision theoretic methods,* in *Computer Based Examinations for Board Certification,* Mancall, et al, Editors. 1996, American Board of Medical Specialties: Everston.

2.Swanson, D., I. Norcini, and L. Grosso, *Assessment of clinical competence: written and computer based simulations.* Assess Eval Higher Educ, 1987. **12**: p. 220-246.

3.Newbie, D., J. Hoare, and A. Baxter, *Patient management problems: issues of validity.* Med Educ, 1982. **16**: p. 137-42.

4.Weinstein, M. and H. Fineberg, *Clinical Decision Analysis.* 1980, Philidelphia: WB Saunders Co.

5.Howard, R. and J. Matheson, *Influence Diagrams,* in *The Principals and Applications of Decision Analysis,* R. Howard and J. Matheson, Editors. 1981, Strategic Decisions Group: Menlo Park, CA. p. 719-62.

6.Shachter, R., *Evaluating influence diagrams.* Operations Research, 1986. **34**(6): p. 871-82.

7.Sox, H., *et al.*, *Measuring the Outcome of Care,* in *Medical Decision Making.* 1988, Butterworth Publishers: Stoneham, MA. p. 167-200.

8.Von Neumann, J. and O. Morgenstern, *Theory of Games and Economic Behavior.* 3rd ed. 1953, Princeton, NJ: Princeton University Press.

9.Howard, R., *Risk Preference,* in *The Principals and Applications of Decision Analysis,* R. Howard and J. Matheson, Editors. 1981, Strategic Decisions Group: Menlo Park, CA. p. 628-63.