Advances in Health Sciences Education **5:** 23–41, 2000.
© 2000 *Kluwer Academic Publishers. Printed in the Netherlands.*

Exploring Alternative Models of Complex Patient Management with Artificial Neural Networks

ADRIAN M. CASILLAS[1,2,*], STEPHEN G. CLYMAN[4], YIHUA V. FAN[4] and RONALD H. STEVENS[1,3]

[1]*Department of Microbiology and Immunology, University of California, Los Angeles School of Medicine, Los Angeles, CA, USA;* [2]*Department of Medicine, Division of Clinical Immunology and Allergy, University of California, Los Angeles School of Medicine, Los Angeles, CA, USA;* [3]*Graduate School of Education and Information Science, CRESST, University of California, Los Angeles, Los Angeles, CA, USA;* [4]*National Board of Medical Examiners, Philadelphia, PA, USA (*Corresponding author: UCLA School of Medicine, Division of Clinical Immunology & Allergy, Los Angeles, CA 90095-1680, USA; E-mail: acasillas@mednet.ucla.edu)*

Abstract. This study applied an unsupervised neural network modeling process to test data of the National Board of Medical Examiners (NBME) Computer-based Clinical Scenarios (CCS) to identify new performance categories and validate this process as a scoring technique. The classifications resulting from this neural network modeling were consistent with the NBME model in that highly rated NMBE performances (ratings of 7 or 8) were clustered together on the neural network output grid. Very low performance ratings appeared to share few common features and were accordingly classified at isolated nodes. This clustering was reproducible across three separately trained networks with greater than 80% agreement in two of the three networks trained. However, the neural network also contained performance clusters where disparate NBME-based ratings ranged from 1 (worst) to 8 (best). Here, agreement between networks was less than 60%. Through visualization of the search strategies (search path mapping), this neural network clustering was found to be sensitive to quantitative and qualitative test selections such as excessive usage of irrelevant tests reflecting broader behavioral classification in some instances. A disparity between NBME ratings and an independent human rating system was detected by the neural network model since disagreement among raters was also reflected by a lack of neural network performance clustering. Agreement between rating systems, however, was correlated with neural network clustering for 92% of the highly rated performances.

Key words: clinical reasoning, medical education, NBME, neural network, problem-based learning

Introduction

Medical diagnosis and management result from the integration of complex behaviors; these behaviors can be decomposed and studied as issues of cost, risk, the quality of life, etc. (Elstein, 1993; Elstein et al., 1982; Fehrsen and Henbest, 1993; Lantz and Chaves, 1997). Currently no stimulus other than the clinical encounter itself adequately captures the information needed to simultaneously address all the aspects of these different patient care models. Some artificial stimuli do exist, however, which are attempting to capture, codify and evaluate the quality of subcomponents of patient care; such stimuli are planned for use in the United States Medical Licensing Examination (USMLE).

For decades, the need for complex performance assessment has been recognized in tests designed for licensure and certification. Early attempts to gain a sense of clinical competence took the form of bedside and oral examinations, which failed primarily due to task and examiner variability (Krome et al., 1983; Norcini et al., 1993). Current testing alternatives have taken the form of "standardized patients" and computer-simulated patient management cases, the latter of which is optimally situated to take advantage of information technology (Clauser et al., 1996). More recently, the National Board of Medical Examiners (NBME) has introduced the computer-based case simulations (CCS) to provide a simulated experience of a patient encounter over time requiring examinees to continually monitor the patient and make appropriate management decisions (Clauser et al., 1991, 1993, 1995, 1996). Over 2300 tests, medications, consultations and procedures are available for the student to choose from in an uncued manner, and each of the actions taken is recorded in a transaction list. These actions are classified based on degrees of appropriateness, inappropriateness and potential risk on a case by case basis. This comprehensive list of actions provides very rich data about student performance, to which appropriate measurement models can be applied to to generate a score. Scores are generated based on a regression model emulating clinicians' ratings of sampled examinee performances on cases (Clauser et al., 1993). Since the main goal of these types of examinations is for licensure, there is a "cutoff" that is ultimately generated in order to derive a minimally acceptable score as a measure of competence.

All forms of scoring are inherently data compression techniques where actions are matched with criteria and compared with broader performance standards. The assignment of a final score may give the impression that overall proficiency or competence can be based on a single criterion derived from the student's model of comprehension. The score is actually derived from multiple variables to ensure optimization of the analytical scoring model for the purposes of the examination (Mislevy and Gitomer, 1996). It must be recognized however, that many aspects of learning, knowledge, and behavior affect performance and not all of these can be addressed in any particular examination format. It is also not always clear what features of problem-solving strategies go unrecognized in this data compression process. Medical licensure assessment is likely to be sensitive to egregious acts of omission or commission, but less significant test usage may be overlooked.

Most performance assessments or "intelligent tutoring" systems begin with an assumed set of knowledge skills and an analysis of cognitive tasks required in order to create suitable testing and scoring criteria (Mislevy and Gitomer, 1996). While the NBME's CCS has been developed and refined for licensure assessment, other more subtle aspects of physician performance may exist within the data that have not been fully explored. These may relate to the more metacognitive aspects of medical problem-solving which deal with task recognition, reflection, uncertainty and deciding when sufficient data to close the case has been gathered (De Bliek et al., 1984). This suggests the possibility for more exploratory analytical techniques that can be readily applied to uncover alternative performance classifications within complex clinical performance data.

The very broad scope of CCS problems allows for alternative classification and assessment models based on performance. The approach we have taken, which we call *constructive data modeling*, utilizes the pattern-recognition capabilities of artificial neural networks to build performance models from existing complex data sets. Artificial neural networks are a non-parametric method of building rich models of complex phenomena through a training and pattern-recognition process and are capable of categorizing behavior based on actual performance sequences. Neural networks have practical utility in solving classification problems with ill-defined categories, where the patterns are often deeply hidden within the data or where there are poorly defined models of behavior (Rumelhart and McClelland, 1986).

Using ANNs, we have previously been able to discern differential aspects of student and clinician diagnostic strategies on infectious disease simulations (Stevens and Lopo, 1994; Stevens et al., 1996). This approach also effectively redefines the generation of performance and scoring models from a model of expert-assigned ratings for ideal performances to one based on actual performances. This approach can be extended to criterion-based (i.e. novice-expert) comparisons as well (Stevens et al., 1996; Stevens and Najafi, 1993).

The purpose of this study was two-fold: first to determine the utility of using artificial neural network analysis to generate performance classifications from complex clinical performance data sets independently of defined scoring criteria, and second, to use neural network analysis to explore the current NBME CCS scoring model.

Methods

THE NBME CCS DATA SET

The NBME developed CCS for assessing physician-patient management skills at the completion of the licensure process (Clauser, 1995, 1996). In these simulations, examinees manage a computer-simulated patient in a realistic fashion by requesting various diagnostic tests, therapeutic options, and other clinically relevant items. The series of requested actions is recorded in simulated time, and the resulting transaction list defines the strategy for each particular student. These sequential actions served as the input to train our neural networks as previously described (Stevens and Lopo, 1994; Stevens et al., 1996).

The transaction data underwent minimal preprocessing to avoid restricting the training data set, which might result in the neural network learning an over-generalization (Elman, 1993). For example, we chose to include all steps of a student's performance whether they were ordered as a package of tests or as individual components to avoid making assumptions about the ordering of data.

Unsupervised neural network analysis of performance data

We took advantage of the ability of neural networks to cluster performance data based upon common features of the examinees' performances. Furthermore, our networks were trained with the data of the performance itself, meaning the sequence of choices taken and not classifications generated *as a result* of the performance. For example, rating criteria such as pass/fail, number of benefits, or number of risks taken by an examinee were deliberately not presented to the network as input that would influence the organization of the performances. These network training conditions are referred to as unsupervised or self-organizing (Kohonen, 1989; Lawrence, 1993) since no "outcome" features of the performance other than the actual steps taken in the process of generating a strategy were part of the input. This is opposed to supervised training where known outcomes (i.e. pass/fail, right/wrong) are associated with the performance in order to classify types of performances with specific outcomes.

The distinct inputs that make up a performance can be simplified to basic units or steps which indicate the transition from one test chosen to the next. This "from-to" pair becomes one element of the total strategy so that the performance becomes a series of "from-to" inputs. The number of inputs is finite: $n^2 - n$, where n is the total number of test items that can be taken and where n is subtracted in the expression to reflect the fact that steps taken in a performance are not between the same choices (i.e. from Test A to Test A). By representing the performance as a series of steps we were able to generate unbiased clusters of performances based on the similarities of the inputs used to generate them. These clusters are identified on an artificial grid or matrix and are given a unique location referred to as a node. These nodes are initially random mathematical representations (vectors), but through the training process, the vectors generated from the input steps in a performance reinforces some of these representations while diminishing others. Eventually, each node represents a general category of performance. A trained neural network presented with a new performance will most closely match the vector generated from that sequence of performance steps with the trained nodes of the network. The output of this analysis is at a node referred to by number.

The number of outputs is a user-specified number and is empirically derived, which for our purposes was 100 outputs arranged in a 10×10 matrix. The self-organizing neural network was constructed with software libraries from Ward Systems Group (Rockville, MD).

The neural network training process is iterative. Each iteration, referred to as an epoch, consists of presenting the entire training data (all the inputs for all the performances) through the network. During each epoch of training, the magnitude and direction of each output is adjusted, and these iterations are continually presented to the network until training is completed and results in consistent output. The total number of epochs or data iterations required to adequately train the network is also empirically derived. Based on our experience, 1000 to 10,000 epochs was sufficient for achieving consistency in performance clustering. This process results in the grouping of performances into clusters, which represent similar performances according to the neural network. (N.B. The number of data iterations (epochs) used in training is not related to the number of outputs or nodes that the data will be clustered into although both values are defined by the user prior to training.)

Following the training process, the same data used to train the network (training set) or novel data (testing set) was presented to the network for classification. The neural network associates each input pattern with a representative output pattern, and the summary of the entire set of performance data produces the topographic representations as shown in Figures 2, 5, and 6.

SEARCH PATH MAP ANALYSIS

To understand the clustering of performances resulting from the ANN classification, we electronically reconstructed the students' problem-solving strategies by generating visualizable strategies or "search path maps". A search path map is generated as a sequential display of the test items chosen by the student while working through a case (Stevens, 1991; Stevens et al., 1989). The tests were grouped to display the use of certain related tests or to show details of the sequence of student selections in a particular test group. In the template used for this study (Figure 1A), the different laboratory tests available were arranged into separate areas representing common tests (i.e. blood tests, antibiotics, etc.). This process is done to simplify the viewing of a strategy as a search path map. The test items, represented as numbered rectangles, can be "selected and dragged" to any location on the computer monitor to facilitate viewing. Individual student performances were overlaid on this template as a series of lines connecting the chosen items in sequence with the lines going from the upper left-hand corner of a test selection to the lower center of the subsequent test (Figure 1B). Where multiple student performances were displayed, the thickness of the lines between test item pairs was proportional to the number of students who made that test selection (Stevens et al., 1996; Stevens et al., 1991)

Results

ARTIFICIAL NEURAL NETWORK CLASSIFICATION OF CASE PERFORMANCE DATA

To determine how simulated clinical performance classifications related to NBME ratings, a training set of 100 randomly selected performances of an emergency infectious disease case was used to train an unsupervised 10 × 10 output node

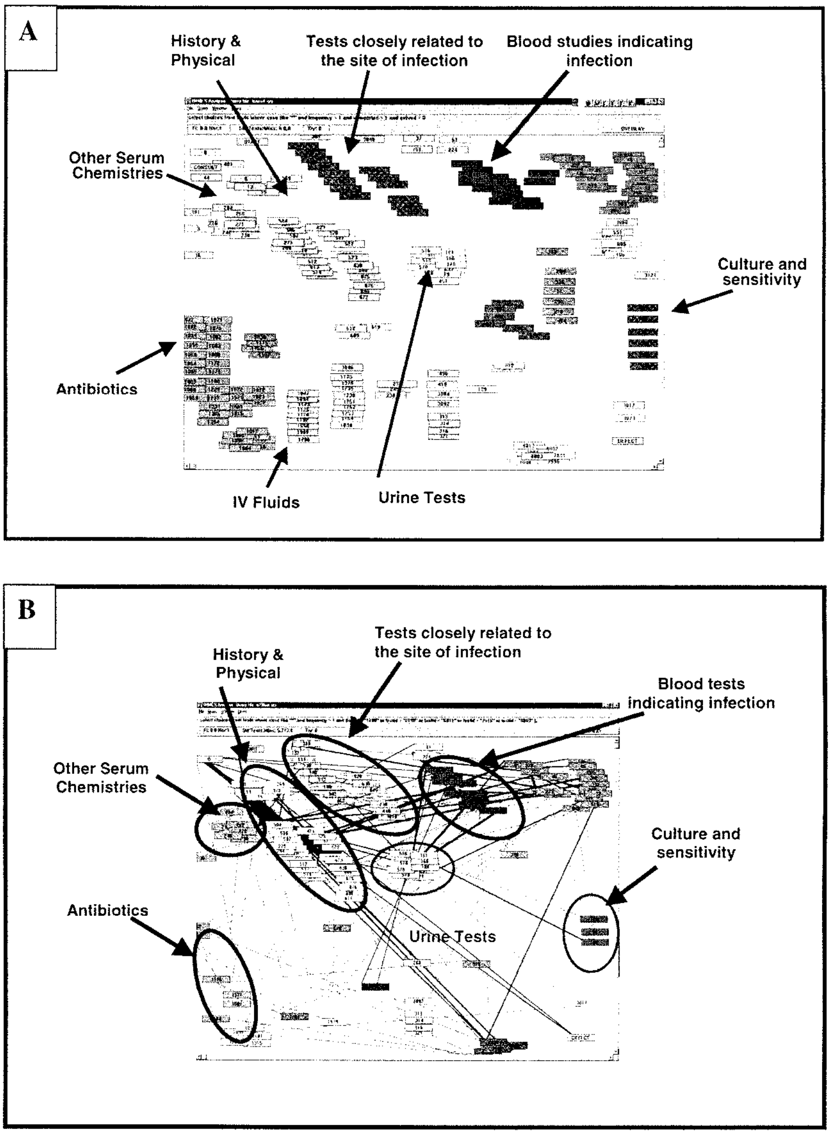

Figure 1. The problem space. A representative template (A) of the test item selections was produced with IMMEX::Analysis as a way to segregate the different types of tests in an area referred to as the problem space. The numbered boxes represent test items that can be logically grouped (i.e. boxes representing tests are selected, dragged, and placed in close proximity to create the various domains shown) by the user of the analysis software according to the type of test or procedure. The labels and arrows have been added to indicate what types of tests the displayed groups represent. Performance strategies for the emergency, infectious disease case can be visualized (B) as a sequence of test items in the order selected by the student (the complexity of this case does, however, make this group of sequences extremely difficult to follow). The thickness of the lines is proportional to the number of times the path between the items was selected (i.e. more students using the path between test selections creates a thicker line).

network for 10,000 epochs. The training data was then analyzed by the same network to identify where the performances clustered on the neural network (Figure 2A). We observed major clusters of student performances at nodes 10 (n = 11), 25 (n = 8), 29 (n = 9), 43 (n = 5), and 44 (n = 7). Since the NBME had previously rated these performances from 1 (worst) to 8 (best), it was possible to determine if any nodes represented clusters with specific ratings. Nodes 43 and 44 were composed exclusively of highly rated performances (rating $\geq$ 7) (Figure 2B), while low ratings (i.e. $<$4) showed limited clustering (Figure 2C), of a few performances per node. In fact, lack of clustering by the ANN was associated with poorly rated performances.

Certain features inherent to a neural network's assignment of clusters, as well as the properties of the data itself can generate variability in classifications across a series of trained neural networks. This variability may lead to differences in classification outputs between networks even though trained with the same data due to the fact that the training process is non-parametric. We were interested to know how well the network-assigned classifications were retained when multiple neural networks were trained with the same data. Two additional neural networks were trained with the same architecture (i.e. 10,000 epochs with a 10 $\times$ 10 output). We selected major nodes from the ANN output of our original network (Network #1) in order to compare these performance clusters on two independently trained neural networks (Networks #2 and #3). The results are summarized in Table I. We anticipated that performances that clustered on one neural network would be represented in a topologically ordered manner (i.e. physically close) on different neural networks, provided the data represent similar patterns (Kohonen, 1989; Lawrence, 1989). We compared six major output nodes generated by Network #1 (nodes 10, 21, 25, 29, 43/44, and 89) to the corresponding outputs generated for the second and third networks. Ninety-one percent of the performances from node 10 of Network 1 clustered similarly on Network #2, while eighty-two percent clustered on Network #3. Between Networks #2 and #3, there was seventy-three percent preservation of clustering for the same performances (node 10, Network #1). Node 21 of Network #1 showed only fifty-percent similarity in clustering with either clustering of the two independent networks; however, Networks #2 and #3 showed identical classification for these performances, since all performance clustering occurred within two nodes. At node 25, there was a 63% similarity between Network #1 and either Networks #2 or #3, but again, Networks #2 and #3 showed identical clustering patterns. The analysis of node 43/44, on Network #1, corresponding to the highest NBME ratings, revealed that sixty-seven percent of these performances also clustered on Network #2, and between Networks #1 and #3, eighty-three percent of the performances clustered at nearby nodes (number 25 and 45) within Network #3. Furthermore, between Networks #2 and #3 there was eighty-three percent clustering similarity of the node 43/44 performances. At node 89, the node representative of the poorest NBME ratings, there was an exact match with the clusters generated by Networks #2 and #3.

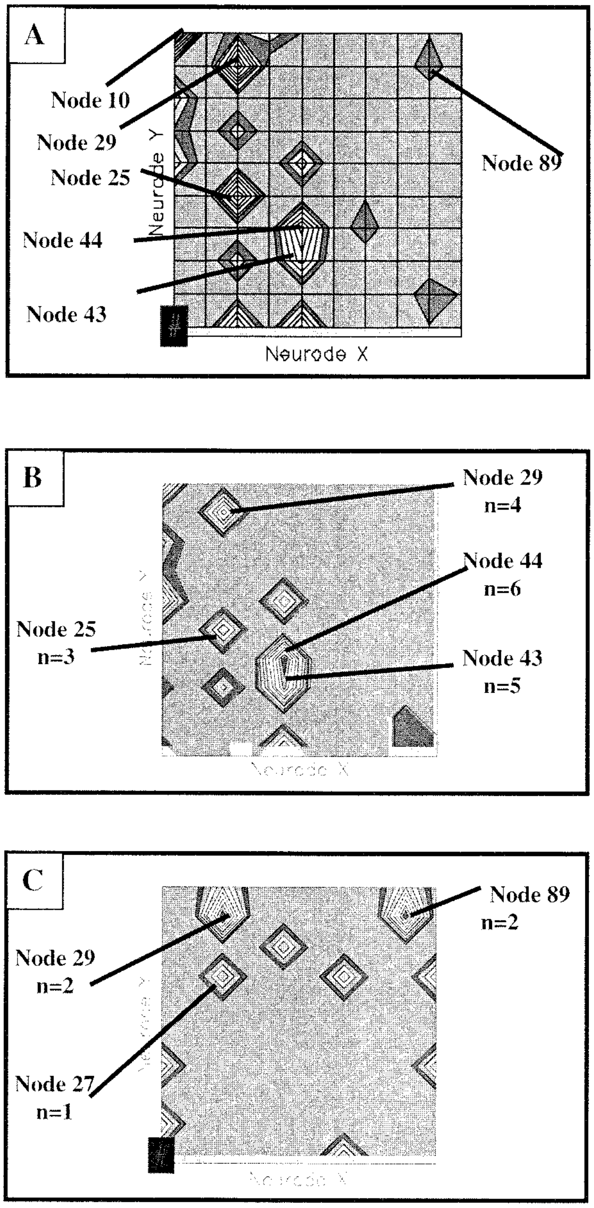

Figure 2. ANN-assigned performance clusters. The 10 × 10 output grid (A) contains a number of clusters of performances which were assigned by the ANN at various nodes for all (n = 100) performances used to train the network. As the ANN analyses the performance, it will assign it to a node. Similar performances will be assigned to the same node or an adjacent one. The more dissimilar a performance is to those assigned to a cluster, the further away it will tend to be assigned. By separating the performances by their NBME-assigned ratings (high ratings in Figure 2B and low ratings in Figure 2C) the nodes that cluster high versus low ratings can be seen. The major nodes are shown with lines pointing to their location on the grid with the number of performances indicated at each node. The grid lines have been removed in 2A and 2B.

Table I. Comparison of performance clustering between specific nodes of Network #1 with two independently trained networks (#2 and #3). All clustering of performances was referenced to the Network #1 population at each node

Network #1 nodes	*n*	Network #1: Network #2	Network #1: Network #3	Network #2: Network #3
10	11	91%	82%	73%
21	6	50%	50%	100%
25	8	63%	63%	100%
29	9	56%	56%	33%
43/44	12	67%	83%	83%
89	2	100%	100%	100%

Although the comparison of performance clusters across a series of networks provided evidence for a structured, non-random grouping of performances, there were instances (such as at node 29) where the similarity in clustering with either of the two independent networks was only fifty-six percent. The node 29 performance clusters on Network #1 were not retained between Networks #2 and #3, since only thirty-three of the performances clustered among these two additional networks. Combined, these data indicated that the reliability of performance clustering was variable across different neural networks for certain performances. This variability suggested that the performance clusters at some nodes represented more complexity than at other nodes.

Understanding the performance-based strategies of the ANN classifications

To better explore the complexity of clustering we observed and to further understand the student strategies assigned to these clusters, representative performances were reconstructed as search path maps on the problem template as described in Methods. Analysis of test usage alone revealed that the degree to which a student searched within a limited set of clinical studies that were closely related to the site of infection was associated with the overall rating (Figure 3). Performance ratings of 7 and 8 (i.e. at nodes 43 and 44) were uniformly associated with the selection of a series of many diagnostic studies associated with the site of infection (Figure 3A) while performance scores below 3 (i.e. node 89) were associated with minimal or no test ordering in that domain (Figure 3B). As expected, the lowest ratings showed no usage in the critical test domain suggesting a failure to recognize the exact source and cause of the infection.

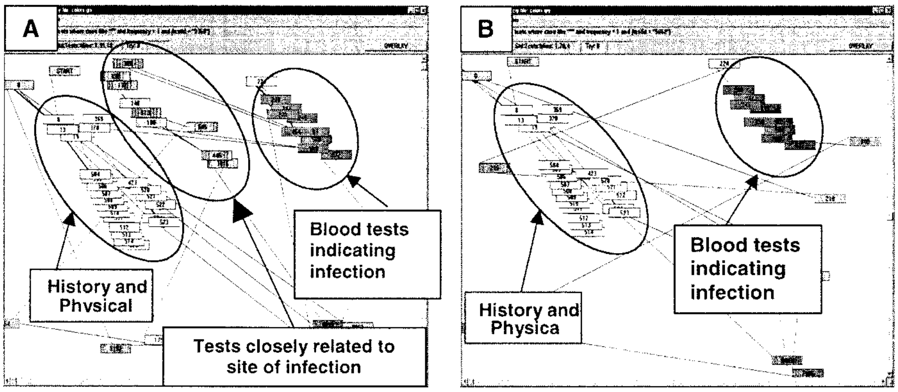

Figure 3. Limited search path maps for performances at node 43/44 and node 89. Search path maps showing a specific group of tests crucial to solving the case are indicated as “tests closely related to the site of infection.” Performances were selected based on where they were clustered by the ANN: those at node 43/44 (A) also had high NBME-assigned ratings (7 or 8) and those at node 89 were rated as failures (B). Note the lack of specific usage of “tests closely related to the site of infection” in (B).

While the neural network consistently clustered some performances in agreement with the NBME-assigned ratings, there were other nodes that contained a range of performance ratings. This finding was characteristic of the cluster at node 29, which incidentally showed the most variability of clustering across networks (Table I). The search path map analysis for this node (Figure 4) indicated the use of excessive test item selections throughout several domains. The consistent feature of these performances was the overuse of tests including serum lipid profiles, urine tests, culture and sensitivity, and additional blood chemistries whether the rating was high (Figure 4A) or low (Figure 4B). By comparison, the highly rated performances at node 43/44 (see Figure 3A) were characterized by a more efficient strategy with a minimal set of test selections by the examinees compared to those at node 29. This data begins to show differences in performance behavior that can be identified for two highly rated yet distinct groups of performances.

Secondary nodal output further defines the quality of performances

The disparate ratings seen at node 29 in this case prompted us to investigate the possibility that there may be additional information within the network output useful for characterizing the quality of these performances. Up to this stage, the network output we had focused on for clustering performances was the node with the highest output or “winning” node (the node at which a performance is ultimately clustered). There are, however, output values assigned to each node within the 10 × 10 output grid for each performance which, if viewed, represent a full topology of the output space. In the next series of experiments, instead of viewing

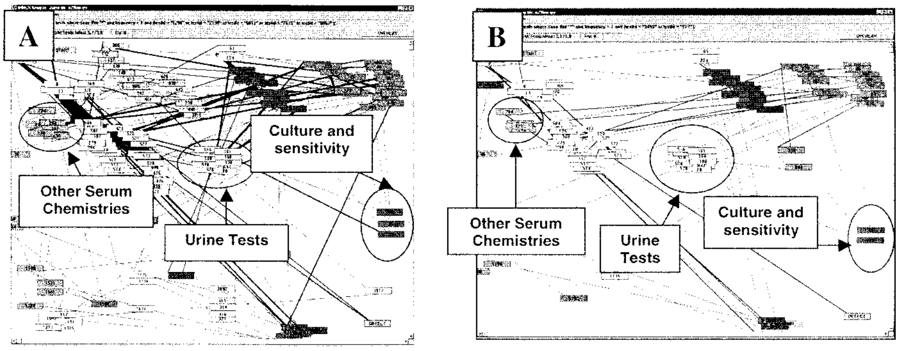

Figure 4. Limited search path maps for performances at node 29. The performances at node 29 are presented by two groups of search path maps: those that were assigned high ratings by the NBME (A) and those with poor ratings (B). A common feature of both types of performances is the excessive ordering of tests by these students. The specific test items are shown for comparison to performances at node 43/44 (Figure 3A) that were solved without the use of the additional tests that characterize these performances.

the outcome of a performance only by its ANN-assigned winning node, all of the outputs generated at the final step of the performance were visualized (Figure 5).

An analysis of the low- and high-rated performances at node 29 was undertaken by visualization of the entire output generated at the last step of the problem solving performance. A representative low-rated performance is shown in Figure 5A. The output generated at the wining node (node 29) was approximately 2400 relative units, and the next highest output area is characterized by a poorly-defined area of output approximating nodes 22/23 and 32/33 (Figure 5A). In contrast, a highly ranked performance (Figure 5B) at the same node (node 29) revealed a clearly defined secondary peak with an output value at node 43/44 only slightly lower than the primary peak. Through the visualization of the entire final output, it is apparent that the high output at node 29 grouped the two performances by associating them with excessive test usage, as was shown previously (Figure 4). However, the development of a significant secondary peak in a defined area associated with an area of high ranking (i.e. node 43/44) served to differentiate the superior strategies from the weaker strategies even at the same node.

Disparity in ANN classification correlates to human rater classifications

As described previously, the NBME develops scores for CCS with regression-based modeling of clinician ratings. There are instances, however, where these two ratings do not agree. We were interested in analyzing the ANN-derived classification of those performances where the NBME score for the case and the clinician ratings of performance for the same case differed by two points on the 1–9 rating scale. In general, we noted that there was agreement between the ratings when

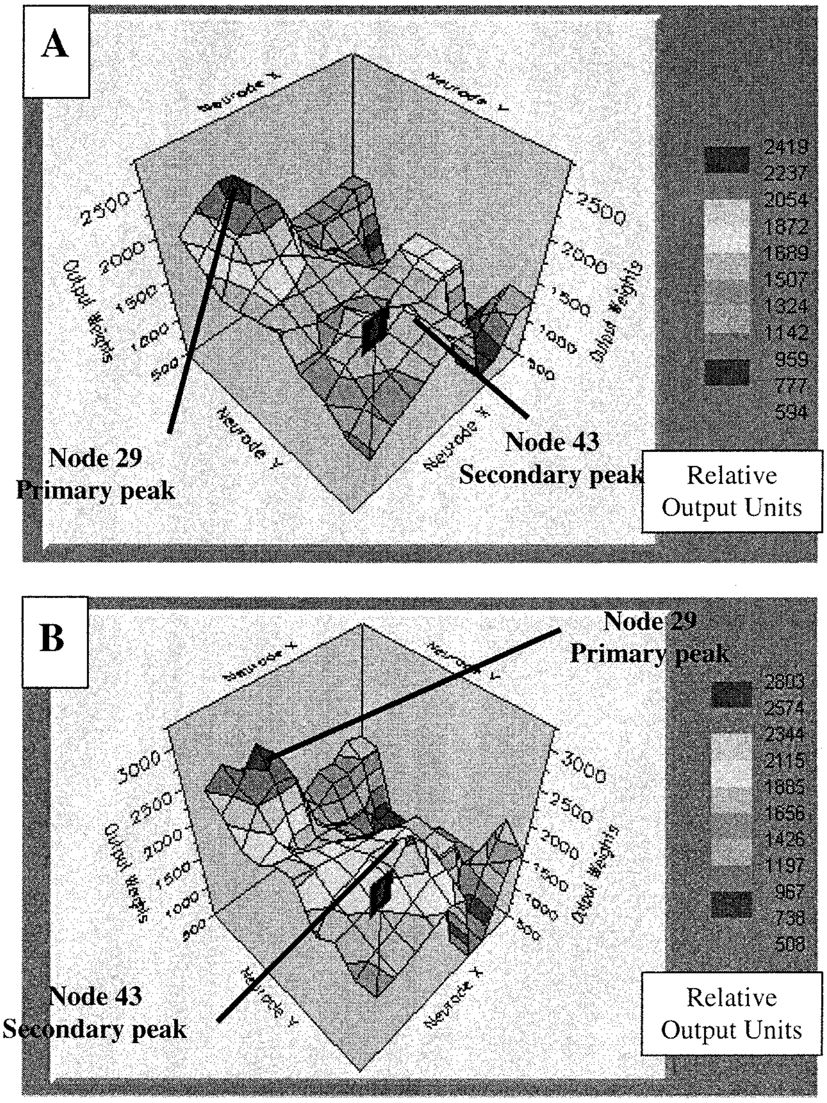

Figure 5. ANN output map of all nodes. All ANN outputs for two representative performances that clustered at node 29 are shown: one with a low NBME rating in (A) and the other with a high rating in (B). The "peaks and valleys" relate to the relative output at each of the 100 nodes; the highest peak value is at node 29 (the winning node) where the performances are clustered. Although the clustering is the same for both peroformances (they have the same primary peak), note the difference in the development of a secondary peak in (B). This secondary peak occurs at node 43 and indicates that this performance contained elements of a typical node 43 performance. This is verified by the fact that this performance was given a high rating. Note the lack of secondary nodal output in (A) compared to the significant secondary peak at node 43 in (B).

Table II. Comparison of NBME and independent rating systems where disparity results in an absolute difference greater than two rating points

Student ID	NBME rating	Independent rating	Difference	Node
6243	6	3.6	2.4	25
4259	6	3.2	2.8	21
3265	6	2.8	3.2	40
2199	7	4.6	2.4	29
3341	7	4.2	2.8	6
9992	7	4.2	2.8	10
3350	8	5.8	2.2	44
6665	8	5.6	2.4	41

Table III. Comparison of NBME and independent rating systems for node 43/44 (ratings are 7 and 8)

Student ID	NBME rating	Independent rating	Difference	Node
1252	7	6.0	1.0	44
1549	7	7.2	0.2	44
2995	8	6.4	1.6	43
3350	8	5.8	2.2	44
3459	8	6.2	1.8	43
3639	7	6.8	0.2	44
4438	8	7.2	0.8	43
4535	7	5.0	2.0	44
6351	8	7.0	1.0	43
8861	8	6.8	1.2	44
9555	8	7.4	0.6	43

the overall rating was low (i.e. <5). Among the seventy-one performances with ratings above 5, eight differed by 2 to 3.2 points (Table II); all others differed by less. The individual nodes representing these performances also indicated this disparity through a lack of clustering. Where the NBME ratings were high, there was also agreement in the ratings. For example, where the two rating systems are least disparate there is a great deal of clustering of performances such as at node 43/44 (Table III). Here, 92% (n = 12) of the performances had NBME ratings of 7 and 8 and independent ratings ranging from 5.0 to 7.4 (Table III).

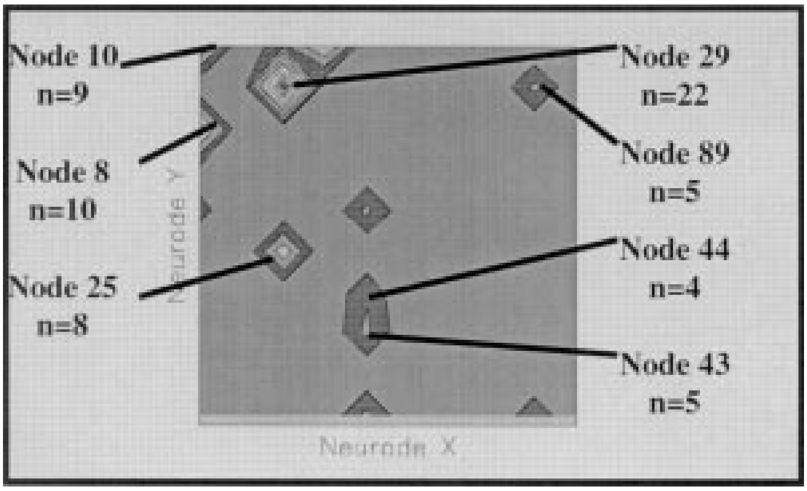

Figure 6. Novel test data output for the infectious disease case. The trained neural network classified novel performances on 100 additional student perfomances. The node and number of performances is indicated.

TESTING THE NETWORK WITH NOVEL PERFORMANCES

One advantage of unsupervised neural network topologies is that the training data can also serve as testing data, a feature important for initial studies given the limited size of the training set (100 performances). We wanted to ascertain how the unsupervised neural network would be able to generalize from the training data to novel performances. We used 100 additional third- and fourth-year medical student performances on the same clinical case. The performances were analyzed by the same method used previously. We observed that significant numbers of novel performances were clustered at the same nodes as we had seen with the training data (Figure 6). This indicated generalizability from the training data to the test data.

The test performances had also been previously rated by the NBME using the same criteria. To determine if the novel performance classifications also agreed with the ratings predicted by the training set, we queried the test data for performances of low and high ratings. Performances from the test data set mainly clustered at nodes 8 (n = 10), 10 (n = 9), 25 (n = 8), 29 (n = 22), 43 (n = 5), 44 (n = 4), and 89 (n = 5) (Figure 6). Performances clustered at Nodes 43 and 44 were associated with high ratings as the training data analysis would predict, but also contained 7 ratings less than 6. Upon analysis of the low ranking performances (i.e. ratings $<$ 4), lack of clustering (i.e. nodes with single performances) was observed in addition to some clustering noted at nodes 29 (n = 22) and 89 (n = 5).

Discussion

The idea of representing models of a cognitive process can be useful when discussing procedures occurring through the acquisition and use of knowledge, and it is this "mental modeling" process that we have addressed in our study. The recurrent performance patterns or strategies that relate external events to what is already known represent these models. Mental models are dynamic, continually being updated and modified with experience in relation to a specific problem. Such models may be rich and lead to an almost automatic retrieval of information when familiar cues are present, as exemplified by medical expertise. They can also be sparse or contain invalid information, leading to uncertainties and perhaps misconceptions. What makes a problem difficult for an individual is often related to how closely a person's mental model of a situation matches the demands of a task or problem. For medical education and assessment, a major challenge is the identification of features of the models that students use as they solve problems and acquire knowledge.

Our approach to generating a neural network-based model of patient management knowledge was deliberately retrospective in approach. The nature of the case and its rating by the NBME were unknown during the neural network training and clustering of student performance. We felt this was important to generate a truly performance-based model that reflected the strategic information and behavioral aspects of problem solving. By utilizing unsupervised neural networks we were able to detect performance patterns that showed consistency between the NBME and neural network which reflected between these two systems. Strong agreement between the neural network and CCS expert raters was observed at node 43/44 (Figure 2B), a high rating by expert raters could be predicted when the output was assigned by the ANN to node 43/44. We have also observed this finding across 3 additional CCS cases (data not shown) where clustering of highly rated performances is a common feature.

We observed that performances with low ratings failed to show clustering (Figure 2C) indicating high variability associated with poor strategies. The poorly rated, divergent strategies reflect elements of inefficiency or error and are expectedly difficult to classify. Compared to traditional testing approaches, it is likely that the CCS format will lead to more random approaches when a mental model is inadequate. This finding is consistent between the CCS expert raters and the ANN models since low ratings reflect poorly cohesive and non-productive strategies indicated by little if any clustering by the ANN (Table II). Moreover, where there is concordance among CCS raters, there is a high degree of ANN performance clustering (Table III).

Further consistency of the NBME model and the neural network model was shown through the training of multiple networks where we noted retention of performance clustering across two additional networks (Table I). These had been trained under the same conditions as the first network, and we found that in four of the five major nodal outputs analyzed, greater than 80% similarity was found in the performances (Table I). This finding could be verified with at least two of the three networks analyzed.

Interestingly, there was low agreement in clustering across the three different networks (less than 60%) for the performances at node 29 on Network #1 (Table I). When the search path maps for these performances were generated, we observed that the network classification was based on an excessive number of tests used (Figure 4). This suggested that among performances characterized by many nonessential tests, there may be strategies, which simply reflect exhaustive searches, as well as those that show the development of a productive approach within, perhaps, a less focused strategy. In either case the element of uncertainty is present, but the model allows for the distinction between these two different outcomes.

Within these strategies clustered at node 29, there was a transition in performance behavior from totally unsuccessful to successful yet diverse approaches. These transitions could be detected in the neural network's output when the total output was viewed for each type of performance (Figure 5). The significance of the development of a distinct secondary output peak at node 43/44 indicated that certain performances contained similarities to excellent performances (as in Figure 3A). This is important when considering the true dynamic nature of learning and comprehension where transitions in performance behavior indicate a learning process.

We made two important observations upon testing novel performances of the case with the trained network. First, the output nodes where novel performances were clustered were similar to the performances from the training data. This proved the utility of the ANN to cluster similar types of performances based on the training data. Second, the testing data reaffirmed the consistencies between the CCS expert-rater model and the neural network model since highly rated performances were again associated with specific nodes (i.e. 43/44) and there was a lack of clustering seen for poorly rated performances (Figure 6). The relevant clusters for the highly rated performances were not as precise, however, as the training data would predict since some performances with overall ratings less than 6 were classified to node 43/44. It is likely that a larger training data set or different ANN architecture will allow for more precise discrimination and less variability in the generalization of novel performances.

Many attempts have been made over the last several decades to capture the cognitive structures utilized by students (McGaghie et al., 1996; Patel et al., 1995). Several forms of assessment have emerged including qualitative cognitive mapping (Stevens, 1991; Stevens et al., 1991) declared strategies through "thinking aloud" protocols (Arocha et al., 1993), multidimensional scaling (de Bliek et al., 1984), and the assessment of semantic structures (Bordage and Lemieux, 1991). Attempts to reconstruct strategies have been most effective in dissecting the knowledge structures used by medical experts versus novices (McGaghie et al., 1996; Stevens and

Lopo, 1994; Stevens et al., 1996) and between groups of medical specialists in order to show transition through the learning process (McGaghie et al., 1994).

The attempts to create models of medical comprehension have attempted to create a standardized clinical assessment environment, which had been difficult to adopt universally for testing purposes (Livingston and Zeiky, 1982; Swanson and Norcini, 1989; Van der Vleuten and Swanson, 1990). With the NBME's CCS project, a scoring algorithm using linear regression techniques to model clinician ratings has provided a more practical means for scoring large-scale performance assessment. Inherent to this method, however, is a mechanistic approach to assessment that may overlook certain unique patterns of behavior that do not fit within an algorithm (Clauser et al., 1995).

ANN-aided assessment may serve to complement more parametric methods where performance behaviors may guide the problem-solving approach. These behaviors reflect the process of objective monitoring of tasks and comparing the newly experienced cues with the knowledge base already formed through similar learned experiences (Brown et al., 1983; Flavell, 1979; Mandin et al., 1997; Patel et al., 1995; Ripley, 1996; Sugrue, 1994). Artificial neural networks could provide a "snapshot" of the performance of a population of examinees, perhaps a particular class at a certain point in time. Learning, as a continuous process, could be viewed as a series of such snapshots revealing individual and group process. We are in the process of pursuing longitudinal studies of problem-solving progress with groups ranging from pre-medical students to highly-trained clinical specialists. As the dynamic nature of learning is modeled, the realistic measures of performance can be tracked for the benefit of students and educators.

Acknowledgements

This work was supported by a Medical Educational Research Grant from the National Board of Medical Examiners to R.S. and a Robert Wood Johnson Foundation Grant to A.C.

References

Arocha, J.F., Patel, V.L. & Patel, Y.C. (1993). Hypothesis generation and the coordination of theory and evidence in novice diagnostic reasoning. *Med. Decis. Making* **13**: 198–211.

Bordage, G. & Lemieux, M. (1991). Semantic structures and diagnostic thinking of experts and novices. *Academic Medicine: Journal of the Association of American Medical Colleges* **66**(9 Suppl): S70–S72.

Brown, A.L., Bransford, J.D., Ferrara, R.A. et al. (1983). Learning, remembering, and understanding. In P.H. Mussen, J.H. Flavell & E.M. Markman (eds.), *Handbook of Child Psychology*, 4th ed., 3rd vol., *Cognitive development*, pp. 77–166. New York: Wiley.

Clauser, B.E., Subhiyah, R.G., Piemme, T.E., Greenberg, L., Clyman, S.G., Ripkey, D. & Nungester, R.J. (1993). Using clinician ratings to model score weights for a computer-based clinical-simulation examination. *Acad. Med.* **68**: S64–S66.

Clauser, B.E., Subhiyah, R.G., Nungester, R.J., Ripkey, D.R., Clyman, S.G. & McKinley, D. (1995). Scoring a performance-based assessment by modeling the judgements of experts. *JEM* **32**(4): 397–415.

Clauser, B.E., Swanson, D.B. & Clyman, S.G. (1996). The generalizability of scores from a performance assessment of physicians' patient management skills. *Acad. Med.* **71**: S109–S111.

De Bliek, R., McGaghie, W.C. & Donohue, J.F. (1984). Representation of clinical case cues: A multidimensional scaling demonstration. *Proc. Ann. Conf. Res. Med. Educ.* **23**: 139–144.

Dunbar, S.B., Koretz, D.M. & Hoover, H.D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education* **4**: 289–303.

Elman, J.L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition* **48**: 71–99.

Elstein, A.S. (1993). Beyond multiple-choice questions and essays: The need for a new way to assess clinical competence. *Acad. Med.* **68**: 244–249.

Elstein, A.S., Rovner, D.R. & Rothert, M.L. (1982). A preclinical course in decision making. *Med. Decis. Making* **2**: 209–216.

Fehrsen, G.S. & Henbest, R.J. (1993). In search of excellence – expanding the patient-centred clinical method – a 3-stage assessment. *Fam. Pract.* **10**: 49–54.

Flavell, J. (1979). Metacognition and cognitive monitoring. *Am. Psych.* **34**(10): 906–911.

Kohonen, T. (1989). *Self Organization and Associative Memory*. Berlin: Springer.

Krome, R.L., Wagner, D.K., Munger, B.S. & Lloyd, J.S. (eds.) (1983). *Standardization in Oral Examinations: Oral Examinations in Medical Specialty Board Certification*. Chicago: American Board of Medical Specialties, 25 p.

Lantz, M.S. & Chaves, J.F. (1997). What should biomedical sciences education in dental schools achieve? *J. Dent. Educ.* **61**: 426–433.

Lawrence, J. (1993). *Introduction to Neural Networks*. Nevada City, CA: California Scientific Software Press.

Livingston, S.A. & Zeiky, M.J. (1982). *Passing Scores*. Princeton, NJ: Educational Testing Service.

Mandin, H., Jones, A., Woloschuk, W. & Harasym, P. (1997). Helping students learn to think like experts when solving clinical problems. *Acad. Med.* **72**: 173–179.

McGaghie, W.C., Boerger, R.L., McCrimmon, D.R. & Ravitch, M.M. (1994). Agreement among medical experts about the structure of concepts in pulmonary physiology. *Acad. Med.* **69**: S78–S80.

McGaghie, W.C., Boerger, R.L., McCrimmon, D.R. & Ravitch, M.M. (1996). Learning pulmonary physiology: Comparison of student and faculty knowledge structures. *Acad. Med.* **71**: S13–S15.

Mislevy, R.J. & Gitomer, D.H. (1996). The role of probability-based inference in an intelligent tutoring system. *User Modeling and User-Adapted Interaction* **5**(3/4): 253–282.

Norcini, J.J., Stillman, P.L., Sutnick, A.I., Regan, M.B., Haley, H.L., Williams, R.G. & Friedman, M. (1993). Scoring and standard setting with standardized patients. *Evaluation & the Health Professions* **16**(3): 322–332.

Patel, V.L., Kaufman, D.R., Arocha, J.A. & Kushniruk, A.W. (1995). Bridging theory and practice: Cognitive science and medical informatics. *Medinfo.* **8**(2): 1278–1282.

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks* 9, Unsupervised methods, pp. 287–326. Cambridge: Cambridge University Press.

Rumelhart, D.E. & McClelland, J.L. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge: MIT Press.

Stevens, R.H. (1991). Search path mapping: A versatile approach for visualizing problem-solving behavior. *Acad. Med.* **66**(9 suppl): S72–S75.

Stevens, R.H., Kwak, A.R. & McCoy, J.M. (1989). Evaluating preclinical medical students by using computer-based problem-solving examinations. *Acad. Med.* **64**: 685–687.

Stevens, R.H. & Lopo, A.C. (1994). Artificial neural network comparison of expert and novice. *Proc. Annu. Symp. Comput. Appl. Med. Care*, pp. 64–68.

Stevens, R.H., Lopo, A.C. & Wang, P. (1996). Artificial neural networks can distinguish novice and expert strategies during complex problem solving. *J. Am. Med. Inform. Assoc.* **3**: 131–138.

Stevens, R.H., McCoy, J.M. & Kwak, A.R. (1991). Solving the problem of how medical students solve problems. *MD Computing* **8**(1): 13–20.

Stevens, R.H. & Najafi, K. (1993). Artificial neural networks as adjuncts for assessing medical students' problem-solving performances on computer-based simulations. *Comp. Biomed. Res.* **26**(2): 172–187.

Sugrue, B. (1994). Specifications for the design of problem-solving assessments in science (CSE Technical Report No. 387). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Swanson, D.B. & Norcini, J.J. (1989). Factors influencing reproducibility of tests using standardized patients. *Teach. Learn. Med.* **1**: 158–166.

Van der Vleuten, C.P.M. & Swanson, D.B. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teach. Learn. Med.* **2**(2): 58–76.