Dynamic Simulation of Medical Diagnosis: Learning in the Medical Decision Making and Learning Environment MEDIC

Cleotilde Gonzalez[1] and Colleen Vrbin[2]

[1] Dynamic Decision Making Laboratory, Carnegie Mellon University
Pittsburgh, Pennsylvania, United States
conzalez@andrew.cmu.edu
[2] Center for Pathology Quality and Healthcare Research, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania, United States
vrbincm@upmc.edu

Abstract. MEDIC is a dynamic decision making simulation incorporating time constraints, multiple and delayed feedback and repeated decisions. This tool was developed to study cognition and dynamic decision making in medical diagnosis. MEDIC allows one to study several crucial facets of complex medical decision making while also being well controlled for experimental purposes. Using MEDIC, there is a correct diagnosis for the patient, which provides both outcome and process measures of good performance. MEDIC also allows us to calculate cue diagnosticity and probability functions over the set of hypotheses that participants are explicitly considering, based on assumptions of local (bounded) rationality. MEDIC has served in a series of studies aimed at understanding learning in dynamic and real-time medical diagnotic situations. In this paper, we outline the tool and highlight results from these preliminary studies which set out to measure learning.

Keywords: Decision Making, Simulation, Medical Diagnosis.

1 Introduction

We accomplish long term goals by making multiple decisions over time. Dynamic decision making (DDM) is about making decisions in an environment that is changing while the decision maker is collecting information about it [1]. Decision makers in dynamic environments make multiple decisions that are intended to reach some goal, and to keep the system under control within a performance range.

Consider a medical dynamic decision making problem. A patient presents symptoms that indicate possible high blood sugar. Tests indicate high blood sugar and low insulin (i.e., hyperglycemia). The physician's goal is to stabilize the patient's health (keep the blood sugar within an acceptable range). The patient can be diagnosed with diabetes (type 1) as the symptoms (cues) develop over time. Once a diagnosis is made a treatment is given, for example, to take insulin. Insulin often takes a moderate amount of time to have an effect in the body.

A. Holzinger (Ed.): USAB 2007, LNCS 4799, pp. 289–302, 2007.

If the amount of insulin is not well calibrated to the state of the body as it is changing, it is possible that the patient would have too much insulin in the body, low blood sugar, and suffer a hypoglycemic crisis. At that point, the solution needs to come quick, to take some sugar by mouth or drink some orange juice, which often have a fast effect on the body. The ideal situation here is to use feedback about the patient's state to keep the system in balance and under control by adding insulin or sugar without over or undershooting. However, we often know that the perception of feedback about the patient's state is inaccurate and the control of the system is often challenging.

Work on the psychology of decision making suggests people have difficulty managing dynamic systems with multiple feedback processes, time delays, nonlinearities, and accumulations [4]. Researchers have found that decision makers remain sub-optimal even with repeated trials, unlimited time, and performance incentives [3, 5, 6, 7]. We believe that more research is necessary to understand the learning process by which individuals improve their decisions after repeated choices in dynamic tasks. Learning is the process that modifies a system to improve, more or less irreversibly, its subsequent performance of the same task or of tasks drawn from the same population [8]. Learning, among other processes (individual differences in cognitive capacity, biases in general reasoning strategies, complexity of dynamic systems), can help explain much of the variance in human performance on dynamic decision tasks. Research in DDM indicates that although individuals may follow very diverse strategies they tend to evolve towards better control policies after an extended number of practice trials [9, 10].

Our research aims at determining how decision makers learn in DDM tasks. In particular, this paper describes a dynamic simulation in a medical context, called MEDIC and presents three behavioral studies to describe how individuals learn using this simulation.

2 MEDIC: A Dynamic Medical Decision Making and Learning Environment

Decision scientists have typically focused on simple and static laboratory tasks. For instance, in the typical laboratory experiment, participants are asked to make likelihood judgments or select among a small number of usually experimenter provided alternatives. Moreover, participants are often provided with one or a sequence of independent choices, where one choice does not influence the next one. Rarely, decision making research has used tools that are representative of the dynamics of the decision making conditions we experience in the real world. Dynamic tools and methods are needed in order to study the dynamics of human behavior. One area of study that can lead to significant improvements in medicine, specifically medical training, is the development of virtual reality simulators for complex medical procedures [11].

We have developed many tools (microworlds, learning environments, management simulators, etc.) for studying DDM [12]. MEDIC is one of those learning environments that was created to study DDM in simulated medical diagnosis.

The development of MEDIC was inspired by Kleinmuntz's [13] paradigm to test the performance of heuristics in a complex dynamic setting. MEDIC includes all the characteristics of a dynamic task described in Kleinmuntz [13]. Using MEDIC we can manipulate the statistical structure of the DDM task in a manner commensurate with Kleinmuntz's analysis: task complexity (i.e., number of diseases and symptoms), disease base rates, time pressure, test diagnosticity, treatment effectiveness, and treatment risk. In addition, we also designed MEDIC to incorporate several dynamic factors not considered by Kleinmuntz, including feedback delays in tests and treatments, dynamic diagnostic cues (cues that have diagnostic patterns unfolding over time), and temporally sensitive symptoms (symptoms that appear at various stages in the progression of the disease). Thus, MEDIC simulates realistic components to a medical diagnosis task, involving multiple feedback loops and possible delays, from the presentation of symptoms to the modification of the patient's health through treatment.

MEDIC has five phases: 1) presentation of symptoms, 2) generation of a diagnosis, 3) test of a diagnosis 4) submission of a treatment and 5) analysis of outcome feedback. The goal in the task is to diagnose and cure patients who are suffering from one disease. The patient is drawn from a population of patients with a configuration of symptoms-diseases and diseases-treatments associations. These configurations are chosen randomly for each patient.

The main measure of performance in MEDIC is the health meter. The health meter is defined as a percent of health, in a scale from 0 to 100. The value of zero indicates death while a value of 100 indicates full recovery. The user of MEDIC can check the status of the patient's health by looking at a graph which continuously monitors the patient's health. MEDIC presents a sequence of patients with an initial health that can vary, but was kept around 50 for the studies reported here. Each patient can be fully cured if the right treatment is applied for the correct disease according to the probabilities defined in the symptoms-diseases and diseases-treatments matrices. MEDIC is a real-time DDM task in the sense that the state of the system deteriorates unless an action is taken. In this case, the patient's health decreases steadily with the passing of time in the simulation.

Initially, a new patient is shown with his or her various contextual data (age, gender and a picture of the patient). Then, participants are presented with probability tables indicating the symptom-disease associations for the patient. According to the probabilities of association of a symptom to a disease, participants decide to conduct tests that would determine the actual presence or absence of a symptom. However, conducting tests for symptoms takes time. Each test is currently configured to return results in 30 simulation minutes. When the test completes, the result of either "present" or "absent" will display, and only then can a second test be issued. Participants are then asked to make their hypothesis of the probability of the different diseases and then they are asked to decide on the most appropriate treatment for the hypothesized disease. This is done by selecting the treatment according to the disease-treatment probability matrix. At the end of each trial (i.e, patient), feedback is provided. Feedback indicates the accuracy of each of the actions taken, such as the real disease the patient suffers from and the belief assigned to that disease by the participant; the correct treatment according to the effectiveness of that treatment for the hypothesized disease. A score structure is defined for participants to associate their actions to their effects. This score is a main source of learning and it is explored in the initial studies as we will explain below.

Although MEDIC allows one to study several crucial facets of complex medical decision making that are often lost in the laboratory, this simulation is also well controlled for experimental purposes. Using MEDIC, we know the correct diagnosis of the patient, which gives us the ability to derive both outcome and process measures of good performance. Although we cannot develop models that prescribe an optimal set of actions, as is the case in many dynamically complex tasks, we can derive Bayesian models that provide benchmarks of "good" behavior. Also, at any point in time (current information state) we can calculate the probability distribution over the diagnostic hypotheses. Moreover, we can assume local (bounded) rationality and calculate cue diagnosticity and probability functions over the set of diagnostic hypotheses the participants are explicitly considering. The studies investigate the learning process in MEDIC from data collected from university students.

3 Three Learning Studies in MEDIC

3.1 Study 1: Learning the Probability Associations

The objective of the first study was to determine whether participants could learn the probabilistic associations of symptoms and diseases in order to make a diagnosis and provide effective treatment for patients in a simulated dynamic medical decision making task.

Methods. The first study was conducted with six graduate and undergraduate students in a research university. They all came to a laboratory where they were trained in MEDIC, and they were asked to diagnose and treat patients for 1 ½ hours.

Participants were presented with a sequence of patients suffering from one of four diseases. The patient and disease associations were selected randomly from one of the four diseases according to the base rates (0.25). Each of the patients in the sequence had a symptom-disease matrix indicating the probabilistic associations between the symptoms and diseases as shown in Table 1. This matrix is seen by participants in the top part of the screen in the MEDIC simulation and it was the same for all the patients in the sequence.

Table 1. Probability matrix with disease-symptom associations used in Study 1

Disease 1	Disease 2	Disease 3	Disease 4	
0.25	0.25	0.25	0.25	Base Rates
0.5	0.5	0.5	0.5	Symptom 1
0.9	0.1	0.5	0.5	Symptom 2
0.9	0.9	0.1	0.1	Symptom 3
0.5	0.5	0.9	0.1	Symptom 4

In this table, diseases 1, 2 3, and 4 presented above each have different associations with the four symptoms. It has been found that participants tend to use positive-testing strategies, suffering from *confirmation bias* [14,15] and *pseudodiagnostic* selection [16]. Thus, an expected behavior is the tendency to issue tests (for symptoms) that have a high likelihood of confirming a hypothesis. However, participants are not restricted in the number or order of tests they can issue. They are allowed to run up to four tests to identify the presence or absence of up to four symptoms.

Participants issue tests (which take time to execute while the patient's health decreases) to determine which symptoms are present. After receiving the test results, the participant adjusts a "belief meter" to reflect his/her assessment of the probability of the disease being present (associations based on the symptom-disease matrix), on a scale of 0 (not present), to 1 (certainly present).

After completing the belief meter for all four diseases, the participant can either conduct more tests or administer treatment. Once a participant has adjusted the belief meters to indicate the likelihood of each disease and is finished testing, effective treatment must be administered according to disease-treatment probability associations defined in another matrix. In this study, the disease-treatment probabilities were fully diagnostic, indicating that one and only one treatment could be effective for each of the possible diseases.

Scoring methods. Human behavior in MEDIC was measured by calculating a score of behavior. We defined three methods of calculating a score.

Scoring Method 1. In the first method, the score included components that rewarded efficient decisions. A cumulative score was presented to each participant based on the following factors: the participant begins with 2000 points for the first patient; if the participant administers an effective treatment, the patient's health is multiplied by 2 (for an ineffective treatment, no points are added); if the patient dies, the participant loses 100 points; each test issued costs 25 points; and points from the belief meter were based on the normalized value assigned on the belief meter compared to the actual value for the correct disease. The equation for scoring method 1 is below.

Score1 = 2000 + (2 * end health if effective treatment OR 0 if ineffective treatment OR -100 if patient dies) – (25 * number of tests performed) – 50 – (50 * (1 – (actual probability of disease – participants normalized, believed probability of that disease) ^2) . (1)

This score replaced the value of 2000 in the equation for the second trial, and so on. The actual cumulative score was plotted and compared with (1) the maximum possible score, which assumed the patient's health did not deteriorate, no tests were performed and the normalized believed probability exactly matched the actual probability for the correct disease; (2) the maximum possible score adjusted for two tests, which are necessary to run to provide an accurate diagnostic assessment; (3) a neutral score that represented the score per trial if the participant made no decisions, meaning that the patient died, no tests were run, and the belief meter was not adjusted; and (4) the minimum possible score, which assumed the patient died, all four tests were run, and the normalized, believed probability for correct disease was 0.

Scoring Method 2. The second performance measure involved a modified score equation which emphasized task accuracy. This method assigned points for correct diagnosis (+/- 100 points), correct treatment (+/- 100 points), proportion of life preserved, most probabilistic informative testing sequence (i.e. testing for the third symptom, and if present test for the second symptom or if absent test for the fourth symptom) (+/- 100 points), difference between actual probability and believed probability for present disease. The equation for scoring method 2 is below:

Score 2 = 500 +/- 100 +/- 100 + percentage health saved +/- 100 - the absolute value of the difference between actual probability of present disease and guessed probability of present disease . (2)

This score replaced the value of 500 in the equation for the second trial, and so on. The actual cumulative score was plotted and compared with (1) the maximum possible score, which assumed correct diagnosis and treatment, no health deterioration, most probabilistic testing sequence, and accurate probabilistic diagnostic assessment for correct disease on belief scale; and (2) the minimum possible score, which assumed incorrect diagnosis and treatment, patient died, any testing pattern other than the most probabilistic sequence, 0 believed probability for correct disease on belief meter.

Scoring Method 3. The third method to calculate the score focused more on the participants' comprehension of the cues as they relate to the probability matrices. This measure of performance incorporated whether or not the participants correctly identified the most probabilistic options by calculating the frequency with which the participants identified the most probable disease even if that wasn't the disease that was present, as well as the frequency with which effective treatment was provided for the guessed disease, even if the guessed disease wasn't the correct or most probable disease. Testing behaviors were also investigated. The goal of this method was to identify whether the participants understood the probability matrices.

In summary, the different scoring methods above would help us best understand the strategies by which the different individuals attempted to save the life of the series of patients they were presented with, and will allow us to understand their behavior.

Results. Results using Score 1 were plotted against neutral, minimum, maximum and maximum adjusted scores. Examples of two individuals demonstrating best and worst performances in this task over the course of the number of patient trials are shown in Figure 2. These two examples show typical behavior in this task according to Score 1. The bolded line represents each participant's actual cumulative score. In total, 33% of participants performed worse than if they had made no decisions at all in the task, similar to Participant 4 below. The other 77% did not perform much better than the neutral score, similar to Participant 6 (Figure 2).

We then fit the participants' behavior per trial to a simpler score calculation, Score 2, hoping this would help reveal why participants performed so poorly in this task. Results from the Score2 are shown graphically in Figure 3 below for the same two participants that were displayed in Figure 2. Figure 3 displays the participant's recalculated cumulative score (bolded line) as compared to the maximum and minimum cumulative scores.

Using this modified score calculation it appears that the overall performance of the participants is better than using the calculation method in Score 1. The participants, on average, were more accurate than efficient in the task.

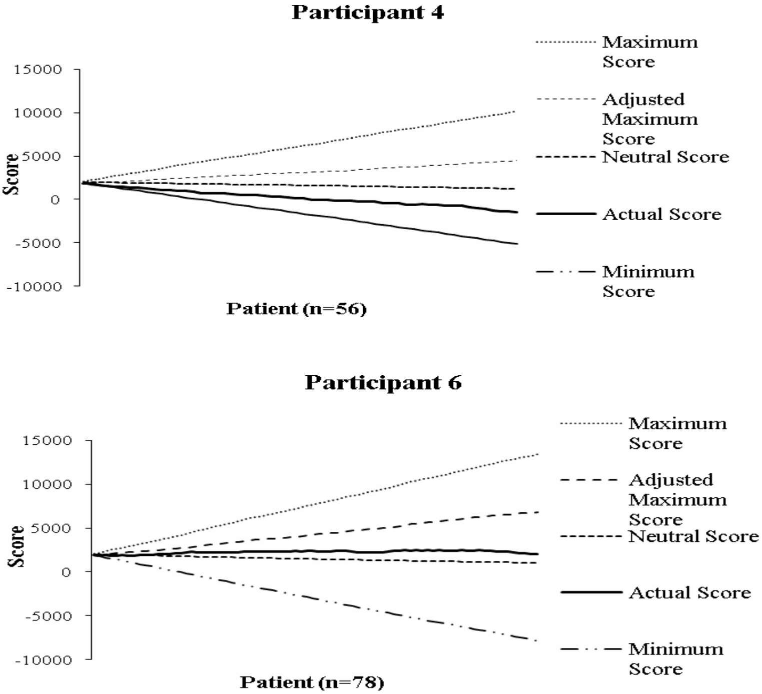

Fig. 2. Best (*participant 6*) and worst (*participant 4*) performances in Study 1 as measured by Score1. The graphs show the maximum score, the adjusted maximum score, the neutral score, the actual score and the minimum score. Participant 6 does not perform much better than the neutral score, and participant 4 performs worse than the neutral score.

A third method for measuring learning was used on the data collected in the first study. This Score 3 helped us identify how participants interpreted the probability matrix. The previous two scoring methods assigned points based on aspects of the task that do not conclusively point to a clear understanding of the probability matrices, such as was the treatment effectiveness and how many tests were run. With scoring method 3, we can identify how well a participant used the cues to identify the most probable disease instead of just adding or subtracting points for accuracy and efficiency.

Table 2 displays the results for Score 3, these are averages across all patients. Since the probabilistic associations between symptoms and diseases have some built in ambiguity (none of the symptoms are 100% associated with any of the diseases), the goal of this method was to determine whether participants could interpret the symptom-disease and treatment-treatment matrices. The results identify a large variability among the six participants in this study. For example, they vary a lot in identifying most probable disease and effective treatment for the chosen disease. Some individuals were very inaccurate in identifying the most probable disease as the real disease. Surprisingly, and despite the fact that the disease-treatment matrix was fully diagnostic, where one and only one treatment was effective for a particular disease, individuals were also ineffective at selecting the treatment with the highest probability of success for their hypothesized, most likely disease.

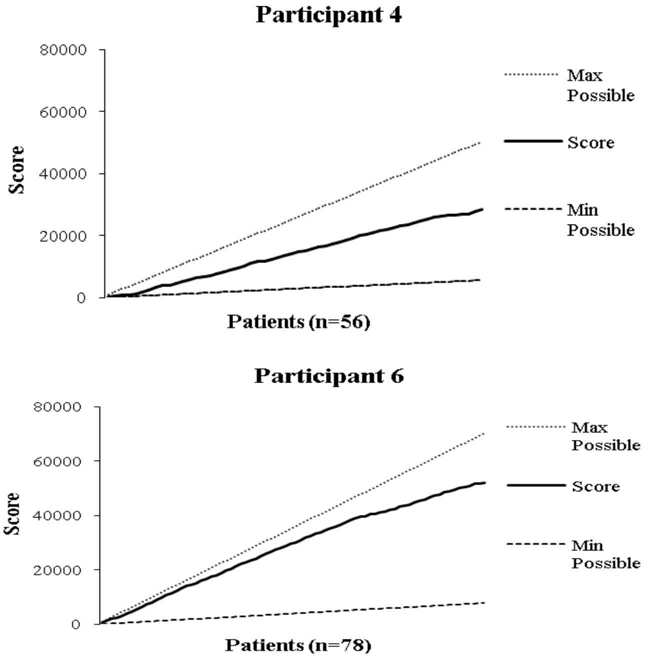

Fig. 3. Performance in Study 1 for two participants measured by Score 2. The graphs show the maximum possible score, the actual score, and the minimum possible score.

Table 2. Performance in Study 1 measured by Score 3 – Diagnosis and treatment behavior

	Indicated Highest Probability on Belief Meter for Most Probable Disease (%)	Chose Effective Treatment for the Disease where Highest Probability was Indicated on Belief Meter (%)
Participant 1	79.2	96.2
Participant 2	89.0	75.6
Participant 3	96.6	94.3
Participant 4	75.0	91.1
Participant 5	64.4	72.9
Participant 6	30.8	33.3

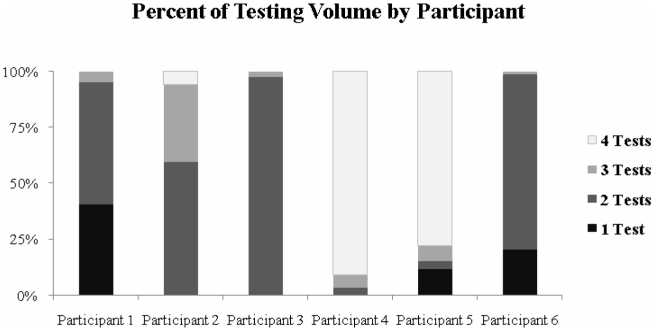

Fig. 4. Performance in Study 1 measured by Score 3

Figure 4 displays the highly variable testing behavior for the six participants. Two of the participants used all 4 tests more than 75% of the trials, whereas two participants used 2 tests for 75% of the trials, and two others used two tests for about half of the trials.

These results are interesting, especially since the test for symptom one does not provide beneficial information, as the symptom is equally associated with all four diseases at 0.5.

Conclusions. We analyzed three different score measures to investigate learning in MEDIC. In the first method, task efficiency was rewarded more heavily. The second method rewarded accuracy. And the final method rewarded comprehension. Performing highly in one of the score calculation methods does not guarantee high performance in another. The different scores allowed us to investigate human behavior at different levels of detail. Most participants did not demonstrate learning, despite the fact that they were allowed extensive practice and despite the full information they were given (both, the symptom-disease probability matrix was given as well as the disease-treatment matrix). The probability matrices were displayed and made fully available to them and they were provided with complete feedback on their behavior. In fact, some of our participants did worse than if they would have made no decisions in the task.

The first score calculation method identified an overall poor performance by all of the participants, who at best performed slightly better than if they had made no decisions in the task. In other words, the decisions that the participants made were not wholly efficient. However, after recalculating the score using the second method, performance appeared much better. This suggests that the participants were able to optimize accuracy better than efficiency in the task. The final score calculation method, which demonstrated probabilistic comprehension, showed that not all of the participants selected the most probable diseases and treatments, suggesting that participants had difficulty in interpreting the probability matrices. Each score calculation method highlights different learning strategies for the task.

Finally, there was notable variability in testing patterns between subjects. This variability inspired the manipulations to MEDIC for the second pilot study, described below.

3.2 Study 2: Learning Probability with Less Ambiguity and Time Constraints

The objective of the second study was to determine whether participants could learn the probabilistic associations of symptoms to diseases in order to make a diagnosis and provide effective treatment for patients in a simulated dynamic medical decision making task, this time without any probabilistic ambiguity and time constraints.

Methods. Most methodological procedures stayed as in Study 1. Thus, here we describe the part of the methods that differed from that first study. Nine participants, graduate and undergraduate students, from a research university were recruited to completed this study. Each participant was asked to diagnose and treat multiple patients for one hour in a modified version of MEDIC; it was modified from the first study to remove the time constraints with both testing delay and decreasing patient health. We also modified the ambiguous associations between symptoms and diseases to include probabilities of either 1 or 0, reducing the ambiguity of the symptom-disease association. The relationships used in this study are shown in Table 3. Given that the Disease-Treatment matrix was fully diagnostic, we removed the treatment section. Thus, the task ended with the participant's diagnosis. The feedback provided included a tally of correct diagnoses and total number of patients seen. According to this new association matrix, the best testing strategy included a total of 2 tests. Participants issuing more than 2 tests would be following a suboptimal strategy, and would be demonstrating a complete lack of understanding of probability relationships.

Table 3. Probability matrix with disease-symptom associations used in Study 2

Disease 1	Disease 2	Disease 3	Disease 4	
0.25	0.25	0.25	0.25	Base Rates
0	0	1	1	Symptom 1
1	0	0	1	Symptom 2
1	1	0	0	Symptom 3
0	1	1	0	Symptom 4

Performance was measured using a score presented to the participant. This score was a tally of correct diagnoses and total number of patients seen.

Results. Table 4 shows each participants performance using the score presented to each participant, which was the total number of correct diagnoses and the total number of trials. Under unambiguous probabilities and no time constraints, all nine participants were able to consistently identify the correct disease.

We then analyzed the testing frequency, as we expected to find a clear and consistent pattern of testing procedures, with a maximum of two tests per participant. Unlike the first study, fewer participants relied on running all four tests, but there was still a considerable amount of variability between subjects who ran 2 versus 3 tests for most of the trials. In this study we also analyzed the range of probabilities individuals indicated for the correct disease when making the diagnosis. Since all ambiguity had been removed from the task, the correct disease was 100% likely to be present. However, 100% was not what participants indicated and surprisingly, ranges varied from .30 to .98. Table 5 displays the results.

Table 4. Performance in Study 2 measured by AVERAGE frequency of correct diagnoses ACROSS TRIALS

	Number of Trials	Correct Diagnosis (%)
Participant 1	244	242 (99.2)
Participant 2	161	161 (100.0)
Participant 3	201	201 (100.0)
Participant 4	211	209 (99.1)
Participant 5	207	206 (99.5)
Participant 6	195	193 (99.0)
Participant 7	170	160 (94.1)
Participant 8	267	259 (97.0)
Participant 9	232	230 (99.1)
Total	1888	1861 (98.6)

Table 5. Pilot Study 2-Measuring learning – Guessed probability for correct disease

	Average Guessed Probability for Correct Disease	Range of Guessed Probability for Correct Disease
Participant 1	0.97	0.91-0.98
Participant 2	0.96	0.82-0.98
Participant 3	0.49	0.30-0.98
Participant 4	0.97	0.97-0.98
Participant 5	0.91	0.80-0.97
Participant 6	0.97	0.88-0.98
Participant 7	0.97	0.97-0.98
Participant 8	0.95	0.55-0.98
Participant 9	0.84	0.69-0.98
Total	0.85	0.30-0.98

Conclusions. Diagnostic accuracy improved from Study1 to Study 2, where the task was over simplified by reducing all the uncertainty in the symptom-disease probabilities and the time constraints. Despite a clear improvement in accuracy with these simplifications, testing patterns were still variable. Individuals were still suboptimal in their testing patterns and in their perception of the probability of the correct disease after testing for symptoms. With these results in mind, modifications were made for a third study described below.

3.3 Study 3: Monetary Incentives

Given that participants are suboptimal learners even in the simplest possible diagnosis task, with unambiguous probabilities and no time constraints, we hypothesized that the only possible explanation left for this performance was motivation. The third study was designed to provide participants with a monetary incentive in the task in order to demonstrate whether participants could learn the probabilistic associations of symptoms to diseases while using the optimal testing strategy and having accurate probability assessments. Participants were assigned to one of two conditions that could earn a bonus. In one condition, a bonus was earned for running two tests, since using the unambiguous probabilities from Study 2 require only two tests to be run in order to make an accurate diagnosis. In the second condition, a bonus was earned by accurately assessing the probabilities of all four diseases each trial.

Methods. This study was identical to the second pilot study, with the exception being the score now represented dollars earned. Nine graduate and undergraduate students were part of this study. The participants were split into 2 conditions. Both conditions incorporated financial incentives of $0.02 per trial to reduce variability while maintaining the level of diagnostic accuracy seen in the second pilot study.

In one condition, which contained 5 participants, a bonus was earned for the ideal testing behavior, which with the unambiguous probability matrix (the same as in Study 2) meant that only two tests were necessary to have complete confidence in a diagnosis.

In the other condition, which contained 4 participants, a bonus was earned for assigning accurate probabilities on the belief scale for all four diseases. With the probability matrix the same as Study 2, the correct disease had a probability of 1, and the other three had a probability of 0.

Each participant was asked to complete 200 trials. Only one of the nine participants was unable to finish, but completed 107 of 200 trials.

Results. Table 6 contains a summary of each participant's performance based on diagnostic accuracy. Although several participants did not perform as well as others, overall the performance was better when individuals earned a bonus for the ideal

Table 6. Performance in Study 3 measured by frequency of correct diagnoses. This table presents the averages across patient trials.

		Number of Trials	Percent of Trials with Correct Diagnosis	Percent of Trials with Bonus Earned
Testing Bonus	Participant 1	200	100%	97%
	Participant 2	200	95%	85%
	Participant 3	200	100%	100%
	Participant 4	200	99%	98%
	Participant 5	200	93%	90%
	Total	1000	97%	94%
Diagnostic Probabilistic Accuracy Bonus	Participant 6	107	70%	0%
	Participant 7	200	100%	93%
	Participant 8	200	100%	95%
	Participant 9	200	57%	0%
	Total	707	83%	53%

testing behavior compared to the participants which earned a bonus for determining the correct probability of the real disease. Surprisingly, 2 out of 4 participants did not earn a bonus for any trial in this second condition.

Both testing behavior and guessed probabilities were analyzed. Providing a financial incentive for optimal testing frequency in this study led to less variability in testing volume. The participants learned that only two tests were necessary, likely by receiving $0.02 cents when the testing strategy was ideal. However, when the financial incentive was not provided for testing strategy, but instead for diagnostic probabilistic accuracy, testing variability was similar to the results seen in Study 2. This suggests that the participants can properly interpret the provided probability matrices, but are willing to run excessive tests in the absence of a financial incentive.

Conclusions. Interestingly, earning a bonus for the ideal testing behavior of two tests greatly reduced the variability of testing behavior between participants earning the same bonus. However, diagnostic probabilistic accuracy did not seem to experience the same decrease in variability when a bonus could be earned.

4 Discussion

MEDIC was developed to study learning in a complex dynamic setting. Many aspects of the simulation can be manipulated, allowing for a variety of experiments. Simulations can be run with or without time constraints, with different levels of feedback, with varying symptoms, tests, diseases and treatments. MEDIC is an important step toward the development of a tool for training and/or reference by medical professionals in decision-making tasks.

The potential applications of MEDIC can be classified into two broad categories: (1) to study cognitive processes underlying physicians' learning of symptoms as they relate to infectious diseases, and, (2) to understand behavior so as to design and implement decision support technology that would assist dynamic decision making under time constraints. Studies that examine cognitive processes focus on understanding the factors that affect hypothesis generation. As described in the studies reported here, decision making using MEDIC can be studied by manipulating the probability matrix that relates symptoms to diseases as well as the types of feedback provided to physicians.

The studies that we ran with MEDIC and that are reported in this paper, demonstrate that even in the simplest possible conditions, with no time constraints and no ambiguity in the symptom-disease probabilities, participants with a high level of education are unable to perform optimally. We also showed that incentives played a key role in their effort and attention they put in to finding the best testing strategy and the determining the appropriate probabilities of the different diseases. The question is then: How can we improve performance in real-world medical diagnosis tasks, where there are immense complexities, time constraints and lack of motivation?

MEDIC allows one to study several crucial facets of complex medical decision-making that are often lost in the laboratory, while also being well controlled for experimental purposes. Using MEDIC, we know the correct diagnosis of the patient, which gives us the ability to derive both outcome and process measures of good performance. Overall, MEDIC provides the necessary paradigm to test the dynamics of hypothesis generation; it also provides data to support the design of medical diagnosis technology that would compensate for deficiencies underlying human cognition under conditions of high workload. We aim at continuing to study the effects of probability uncertainty, time constraints, and feedback on medical diagnosis, and we think MEDIC will support this goal.

Acknowledgments. This research was partially supported by the National Science Foundation (Human and Social Dynamics: Decision, Risk, and Uncertainty, Award number: 0624228) award to Cleotilde Gonzalez.

References

1. Edwards, W.: Dynamic Decision Theory and Probabilistic Information Processing. Human Factors 4, 59–73 (1962)
2. Brehmer, B.: Strategies in Real-Time, Dynamic Decision Making. In: Hogarth, R.M. (ed.) Insights in Decision Making, pp. 262–279. University of Chicago Press, Chicago (1990)
3. Sterman, J.D.: Misperceptions of Feedback in Dynamic Decision Making. Organizational Behavior and Human Decision Processes 43(3), 301–335 (1989)
4. Sterman, J.D.: Business Dynamics: Systems Thinking and Modeling for a Complex World. McGraw-Hill, Boston (2000)
5. Diehl, E., Sterman, J.D.: Effects of Feedback Complexity on Dynamic Decision Making. Organizational Behavior and Human Decision Processes 62(2), 198–215 (1995)
6. Sterman, J.D.: Modeling Managerial Behavior: Misperceptions of Feedback in a Dynamic Decision Making Experiment. Management Science 35(3), 321–339 (1989)
7. Sterman, J.D.: Learning in and about Complex Systems. Systems Dynamics Review 10, 291–330 (1994)
8. Simon, H.A., Langley, P.: The Central Role of Learning in Cognition. In: Simon, H.A. (ed.) Models of Thought, vol. II, pp. 102–184. Yale University Press, New Haven London (1981)
9. Gonzalez, C.: Learning to Maker Decisions in Dynamic Environments: Effects of Time Constraints and Cognitive Abilities. Human Factors 46(3), 449–460 (2004)
10. Kerstholt, J.H., Raaijmakers, J.G.W.: Decision Making in Dynamic Task Environments. In: Ranyard, R., Crozier, W.R., Svenson, O. (eds.) Decision Making: Cognitive Models and Explanations, Routledge, London, pp. 205–217 (1997)
11. Scerbo, M.W.: Medical Virtual Reality Simulators: Have We Missed an Opportunity? Human Factors & Ergonomics Society Bulletin 48(5), 1–3 (2005)
12. Gonzalez, C., Vanyukov, P., Martin, M.K.: The Use of Microworlds to Study Dynamic Decision Making. Computers in Human Behavior 21(2), 273–286 (2005)
13. Kleinmuntz, D.N.: Cognitive Heuristics and Feedback in a Dynamic Decision Environment. Management Science 31(6), 680–702 (1985)
14. Wason, P.C.: Reasoning. In: Foss, B.M. (ed.) New Horizons in Psychology, Penguin, Baltimore, pp. 135–151 (1966)
15. Wason, P.C.: Reasoning about a Rule. Quarterly Journal of Experimental Psychology 20, 273–281 (1968)
16. Doherty, M.E., Mynatt, C.R., Tweney, R.D., Schiavo, M.D.: Pseudodiagnosticity. Acta Psychologica. 43, 111–121 (1979)