COMET: A Collaborative Tutoring System for Medical Problem-Based Learning

Siriwan Suebnukarn, Thammasat University Dental School

Peter Haddawy, Asian Institute of Technology

Clinical reasoning is the cognitive process by which physicians synthesize clinical case information, integrate it with their knowledge and experience, and diagnose or manage the patient’s problem.1 Problem-based learning is a method often used to characterize medical problems as a context for students to acquire knowledge and clinical-reasoning skills. PBL instructional models vary, but the general approach involves student-centered, small-group, collaborative problem-solving activities. The main arguments for using collaborative problem solving in medical PBL include the wider range of ideas generated and the higher quality of discussion that ensues. In addition, students obtain training in consultation skills and group clinical problem solving, which are important for a successful clinical medicine practice.

However, effectively implementing PBL in the clinical curriculum is difficult. First, no standards exist for PBL tutoring,2 and properly trained tutors are lacking. In addition, effective PBL requires the tutor to give students a high degree of personal attention. In the current academic environment, where resources are becoming scarcer and costs must be reduced, providing such attention isn’t necessarily feasible. Consequently, medical students often don’t get as much facilitated PBL training as they might need or want.

Intelligent technologies can enrich medical training by providing environments that maximize learning while minimizing patient risks. Most research on intelligent medical-training systems has focused on particular domains—for example, pathology to train students in feature perception and disease classification.3 Little or no research has addressed a general domain-independent framework for intelligent medical tutoring, and none has addressed intelligent medical tutoring in group settings.

We’ve developed a collaborative intelligent tutoring system for medical PBL called COMET (Collaborative Medical Tutor). COMET uses Bayesian networks to model the knowledge and activity of individual students as well as small groups. It applies generic tutoring algorithms to these models and generates tutorial hints that guide problem solving. An early laboratory study shows a high degree of agreement between the hints generated by COMET and those of experienced human tutors.4 Evaluations of COMET’s clinical-reasoning model and the group reasoning path provide encouraging support for the general framework.5

Medical problem-based learning

PBL is designed to challenge learners to build up their knowledge and develop effective clinical-reasoning skills around practical patient problems.1 It’s typically carried out in three phases:

• Problem analysis: In group discussion, students
evaluate the patient problem presented to them exactly as they would a real patient, attempting to determine the possible underlying anatomical, physiological, or biochemical dysfunctions and to enumerate all possible causal paths (hypotheses and their causal relations) that would explain the progression of the patient’s problems.

- **Self-directed study:** Students work individually outside the tutorial session, addressing any open issues identified in the first phase through relevant learning resources such as literature, laboratories, and specialists.
- **Synthesis and application of newly acquired information:** Students analyze data and wrap up the problem collaboratively after they return from their self-study period.

One main issue in PBL is the tutor’s role. Like a good coach, a tutor needs enough command of whatever the learners are working on to recognize when and where they most need help. The ideal tutor should be an expert in both learning content and learning process, which is rare among human tutors. The tutor intervenes as little as possible, giving hints only when the group appears stuck or off track. In this way, the tutor avoids making the students dependent on him or her for their learning.

**A collaborative medical tutor**

**COMET** aims to provide an experience that captures the tutor’s essential role in facilitating PBL groups. The current version supports the initial PBL problem analysis phase, where the interactions are intensive between students and the system as well as among the students in the group.

We’ve implemented the system as a Java client/server combination that works over the Internet or local area networks and supports any number of simultaneous PBL groups. Students in a COMET PBL group can be in disparate locations but share the same environment over the network. There’s no technical limit on the number of students in a COMET PBL group, but for practical pedagogical reasons, a group typically consists of no more than eight students.

Figure 1 shows the basic system components. The four primary components are similar to any typical intelligent tutoring system domain clinical-reasoning model (or domain model), student clinical-reasoning model (or student model), pedagogic module, and interface, which includes the student multimodal interface and the medical-concept repository.

Our prototype incorporates substantial domain knowledge about head injuries, stroke, and heart attack. These domains are quite different. The knowledge used to reason about head injury is primarily anatomical, while that used to reason about strokes and heart attacks is primarily physiological. Furthermore, the pathophysiology of the latter two diseases is more dynamic. The system implementation is modular, and the tutoring algorithms are generic. So, adding a new scenario requires only adding the appropriate model representing how to solve a particular case, such as a peptic-ulcer domain model. The student model is a probabilistic overlay of the domain model. **COMET** constructs it at runtime by instantiating the nodes that define each student’s knowledge and activity.

**Multimodal interface**

Live human-tutored medical PBL sessions are facilitated by the use of a whiteboard or similar device during problem analysis and hypothesis development. Participants can use the whiteboard to edit and expand ideas. The students are encouraged to express all their ideas in this process and to challenge each other’s ideas as they proceed under the tutor’s guidance. The students use the knowledge and skills they already have, aided by handy references such as a medical dictionary and a few appropriate textbooks on anatomy, physiology, and the like. **COMET** emulates this PBL environment by incorporating a multimodal interface (see figure 2), which integrates text and graphics in a rich communication chan-
COMET’s medical-concept repository helps point to on it.

and anything that other students sketch or rent focus. All students see the same images relevant to the group discussion’s curriculum. COMET displays system input. However, the capabilities, so it can’t use the chat pane text as pertaining into the chat pane (figure 2, middle left pane). COMET has no language processing capabilities, so it can’t use the chat pane text as system input. However, the discussion pane (figure 2, upper left pane) supports collaborative interaction by displaying tutoring hints and student chat dialogues. In the image pane (figure 2, upper right pane), COMET displays images relevant to the group discussion’s current focus. All students see the same image and anything that other students sketch or point to on it.

Along with facilitating system input, COMET’s medical-concept repository helps students better understand the relationships between domain concepts. The repository stores all valid and invalid hypotheses relevant to all problem scenarios, so the system doesn’t need to process typed text input.

Domain and student clinical-reasoning model

Generating appropriate tutorial actions requires a model of the students’ understanding of the problem domain and of the problem solution. However, as in human-tutored PBL sessions, COMET must provide an unrestricted interaction style that gives students the freedom to solve the patient case without having to explain the rationale behind their actions. This complicates the modeling task because the system must infer the student’s level of understanding from a limited number of observations. To deal with the resulting uncertainty, we selected Bayesian networks as our modeling technique. To construct the model, we used information from various sources. We obtained the generic network’s structure from research articles and textbooks on clinical reasoning and PBL. For each problem scenario, we consulted medical textbooks and expert PBL tutors (details on the network’s structure are available elsewhere).

The structure contains two types of information:

• the problem solution’s hypotheses and causal links (see figure 3, right half) and
• how students derive the hypotheses (see figure 3, left half).

We represent the hypothesis structure following Paul Feltovich and Howard Barrows’ model, which defines three illness feature categories: enabling conditions, faults, and consequences. The right half of figure 3 shows five possible faults associated with the single consequence, chest pain: myocardial infarction, angina, musculoskeletal injury, gastrointestinal disorder, and stress. Atherosclerosis is the enabling condition of Myocardial infarction and Angina. The remaining hypothesis nodes are consequences of myocardial infarction. Each hypothesis node has parent nodes, which have a direct causal impact on it. For example, right heart failure has parents pulmonary congestion and myocardial infarction. All hypothesis nodes have two states, indicating whether the student knows that the hypothesis is valid for the case. For every hypothesis that directly causes another hypothesis (for example, atherosclerosis and myocardial infarction), a node (for example, Link_14) represents the causal link between them. The two hypothesis nodes (atherosclerosis, myocardial infarction) are that link node’s parents. The intuition is that the student can’t create the link unless someone has first created both hypotheses. Each link node has two states, indicating whether the student creates a causal link between two hypotheses.

The derivation of hypotheses (figure 3, left half) is represented in terms of three kinds of nodes: goals, general medical knowledge, and apply actions. Every hypothesis node has a unique Apply node as one of its parents. The Apply node represents a medical concept applied to a goal to derive the hypothesis. Part of the modeling process is to come up with a discrete medical concept that underlies a unique hypothesis. For example, the Apply_13 node indicates that the student can use knowledge of the vessel lumina occlusion medical concept to infer that myocardial infarction is a consequence of atherosclerosis. Each hypothesis node thus has a conditional-probability table specifying the probability of the hypothesis being known conditioned on whether the parent hypotheses are known and whether the student can apply the appropriate piece of knowledge to determine the cause-effect relationship. The conditional-probability tables for the Apply nodes are simple AND gates.

We obtained each resulting network’s conditional-probability tables by learning from PBL session transcript data. The data for this study consisted of tutorial session tape recordings and photographs for head injury, stroke, and heart attack scenarios at Thammasat University Medical School. The study included 15 groups of third-year medical students. Each group consisted of eight students with different backgrounds. We presented each
we would like to identify a path that has much of the whole group’s attention and at least one member’s focused attention. COMET uses an algorithm that identifies what we call the group path by determining the probabilities of the various candidate paths, given the hypotheses the students have created thus far.5

**Generating tutorial hints**

Our automated tutor guides the tutorial group to construct possible case hypotheses by asking specific open-ended questions. It aims to give hints when the group appears to be stuck, off track, collaborating poorly, or producing erroneous hypotheses. To do this, the tutor requires knowledge of both the problem domain and the problem-solving process. From our study of the tutoring-session transcripts, we identified and implemented seven hint strategies that human tutors commonly use:

1. **Focus group discussion.** Group members might suggest various valid hypotheses without focusing on any given causal path. When such lack of focus becomes apparent, COMET intervenes by directing the students to focus on one hypothesis in the group path.

2. **Promote open discussion.** If a student proposes a hypothesis that’s not on the current group reasoning path, COMET provides positive feedback by encouraging the student to relate the hypothesis to the current discussion’s focus.

3. **Deflect uneducated guessing.** If a student creates an incorrect causal link, COMET points this out and encourages the student to correct the error.

4. **Avoid jumping critical steps.** If a student creates a link that jumps directly from one hypothesis to a downstream consequence, leaving out intermediate hypotheses, COMET asks the student for the more direct consequences.

5. **Address incomplete information.** After the students have finished elaborating all hypotheses on the group path, COMET identifies another path for them to work on.

6. **Refer to experts in the group.** If the students still don’t respond correctly after COMET provides a general and then a more specific hint, the system determines the student most likely to know the answer and refers directly to him or her.

7. **Promote collaborative discussion.** If one student dominates the discussion, COMET asks for input from other students. If a student doesn’t contribute after a certain number of hypotheses have been mentioned, COMET solicits input from that student.

We used the interaction log and both the structure and the probabilities of the Bayesian-network models as input to the algorithms we developed to generate each hint type. All strategies except strategies 3 and 7 use both the structure and the probabilities of the Bayesian network models. Strategy 3 uses only the model’s structure, while strategy 7 uses only a count of the number of inputs from each student. Strategies 1, 2, and 5 use...
the group path discussed earlier. Strategies 1–5 have general and specific versions. COMET first gives a general hint using the parent goal node of the hypothesis it has determined the students should focus on. If the students either don’t respond or respond incorrectly, the system uses the more specific parent medical-concept node. If the students still can’t come up with the correct hypothesis, COMET refers directly to the student in the group most likely to know the answer. If this doesn’t work, the system identifies this topic as a learning objective for study outside the session (tutoring strategy details appear elsewhere).

Figure 4 presents a transcript showing the interaction with the system after the students read the heart attack problem scenario. The system selects hint strategies and content on the basis of student inputs to the hypothesis board.

**Evaluation**

COMET has been tested empirically in both the laboratory and the field. In an early laboratory study, we compared the tutoring hints generated by COMET with those of experienced human tutors. On average, 74.17 percent of the human tutors used the same hint as COMET. The most similar human tutor agreed with COMET 83 percent of the time, and the least similar tutor agreed 62 percent of the time. Our results show that COMET’s hints agree with the hints of the majority of human tutors with a high degree of statistical agreement (kappa index = 0.773).

We also compared the focus of group activity chosen by COMET with that chosen by human tutors. COMET agrees with the majority of human tutors with a high degree of statistical agreement (kappa index = 0.823).

We tested the modeling approach’s validity with student models in the areas of head injury, stroke, and heart attack. Recovering a student’s medical knowledge to students by comparing student clinical-reasoning gains obtained using COMET versus those obtained from human-tutored PBL sessions.

**Study design**

To evaluate the system’s overall impact on student learning, we designed a study to test the hypothesis that a COMET tutorial will result in student clinical-reasoning gains similar to those obtained from a session with an experienced human PBL tutor. We recruited 36 second-year medical students from Thammasat University Medical School—that is, they didn’t yet have PBL experience in stroke and heart attack. We applied stratified random sampling to divide the students into six groups according to their background knowledge. We compared three student groups tutored by COMET with three other groups tutored by experienced human tutors (see figure 5). Initial training in the use of COMET required about 15 minutes. Three medical PBL tutors participated in the study. Rather than providing answers or explanations, PBL tutors intervene actively and opportunistically, posing open-ended questions and giving hints only when the group appears to be getting stuck or off track. The tutors had at least five years’ experience in conducting the Stroke and Heart Attack course at Thammasat.

The study had a pretest/posttest control group design. After the pretest session, students participated in a two-hour problem-solving session. We asked each group to enumerate a network of hypotheses and causal links collaboratively explaining the stroke and heart-attack scenarios. The human-tutored groups created the network on a classroom whiteboard. We counted all valid hypotheses and links created in each group. A stroke and heart attack expert verified the validity of hypotheses and links. We assessed all students on their clinical reasoning before and after the PBL tutorial session to determine each student’s reasoning gains. Then we compared the clinical-reasoning gains between the two groups.

We used the Clinical-Reasoning Problem approach to assess clinical reasoning. Each CRP consisted of a clinical scenario that a specialist physician vetted for clinical accuracy and realism. The pretest used two heart attack cases and two stroke cases to measure each student’s initial ability to solve the problems. Two different posttest cases in each area measured their ability to generalize the clinical reasoning acquired from the PBL tutorial session to new related cases. We asked participants to name the two diagnoses they considered most likely. We also asked them to list the case features they regarded as most important in formulating their diagnoses, to indicate whether these features were positively or negatively predictive, and to give each one a weight. Figure 6 presents an example CRP student result.

To establish reference scores, we asked 10 volunteer general practitioners to complete both sets of CRPs. We selected GPs because they have experience with a broad range of undifferentiated clinical presentations that encompass all medical practice areas. In this respect, they provide the most appropriate standard against which to compare medical students who graduate with a sound background in general medicine but without any specialist knowledge.

**Results**

The GP results showed no statistically significant differences between pretest and posttest scores, indicating that the tests were of approximately equal difficulty (see table 1). The GP scores varied from 88.20 to 91.50, indicating that the questions weren’t trivial.
We used Cronbach’s alpha to measure test reliability. For the students’ pretest and posttest scores, Cronbach’s alpha was 0.901 and 0.921, respectively. A reliability coefficient of 0.80 or higher is commonly considered acceptable.

We used the Wilcoxon test for nonparametric data and matched pairs to examine the differences between pretest and posttest scores on the same group. We used the Mann-Whitney test for unmatched data to detect any differences between COMET and human-tutored groups. Table 2 shows almost identical pretest mean scores for the COMET and human-tutored students, and the Mann-Whitney test confirms no significant differences between COMET and human-tutored groups. The posttest mean scores were significantly higher than the pretest scores for both COMET and human-tutored students (Wilcoxon, p = 0.000), indicating that significant learning occurred. But the average posttest score for the COMET students (65.12) was significantly higher than that for the human-tutored students (59.46) (Mann-Whitney, p = 0.011), indicating that students were learning more in the COMET sessions than in the human-tutored sessions. Comparing the number of hypotheses and links created by each group shows that, on average, the COMET groups created more hypotheses and links (stroke: hypotheses = 45.33, links = 48.33; heart attack: hypotheses = 40.67, links = 44.67) compared to the human-tutored groups (stroke: hypotheses = 35, links = 40; heart attack: hypotheses = 33.33, links = 38).

We didn’t expect the clinical-reasoning gains for COMET-tutored students to be higher than those for human-tutored students. This is particularly true in light of our earlier study showing that, on average, 74 percent of human tutors used the same hint strategy and content as COMET.1 We believe the explanation lies primarily in the 26 percent disagreement. Human tutors often give up after providing a general hint, jumping right to identifying the hypothesis as a learning objective. In contrast, COMET is more relentless in pushing the students, always following the sequence of general hint, specific hint, referral to an expert, and finally identifying a learning objective. It’s generally agreed that students should generate as many hypotheses as possible in a PBL session, leaving only the truly difficult issues as learning objectives.

**Discussion**

Most cognitive tutors, such as those developed by Cristina Comat, Abigail Gertner, and Kurt VanLehn3 and by Rebecca Crowley and Olga Medvedeva,9 produce hints when students request help and bug messages when they err. Every entry a student makes in the problem-solving interface receives immediate feedback, whether it’s correct or incorrect—for example, green and red in ANDES.9 Following PBL tutorial principles, however, students generally don’t ask the tutor when they get stuck, and the tutor doesn’t say whether the students’ idea is right or wrong. COMET therefore has no button for the student to ask for help and doesn’t indicate whether the student’s entry is correct or incorrect. To recognize when and where the group needs help and to give hints that will help them continue their discussion, COMET uses its strategies to
focus group discussion (strategy 1) and to promote collaborative discussion (strategy 7); these strategies help the group continue its discussion on a productive track. The strategies to avoid jumping the critical steps (strategy 4) and deflect uneducated guessing (strategy 3) are forms of bug messages.

Most strategies have general and specific versions. In COMET, the general hint uses the parent goal node of the hypothesis that the students should focus on; the more specific hint—triggered when there’s no student response or an incorrect response—uses the more specific parent medical-concept node. Generating general and specific hints from ANDES. In ANDES, students who don’t know what to do after receiving the general hint can select a follow-up question by clicking on the three buttons:

- “Explain further” gives slightly more specific information about the proposition represented by the node;
- “How do I do that?” finds the lowest probability child node, assuming that it’s the node the student is most likely to be stuck on; and
- “Why?” displays a canned description of the rule used to derive the node.

In COMET, the system instead gives a specific hint based on the medical concept of the node for the hypothesis the student is most likely to know. If the student still can’t come up with the hypothesis of interest, the system refers the student to experts in the group (strategy 6). As we described earlier, if this doesn’t work, COMET identifies this topic as a learning objective for study outside the session.

Table 1. Mean scores for all Clinical-Reasoning Problems.*

<table>
<thead>
<tr>
<th>CRPs</th>
<th>General practitioner’s score (standard deviation)</th>
<th>Student’s score (standard deviation)</th>
<th>COmET</th>
<th>Human tutor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 10</td>
<td>n = 18</td>
<td>n = 18</td>
<td>n = 18</td>
</tr>
<tr>
<td>Pretest</td>
<td>1.1</td>
<td>88.70 (0.78)</td>
<td>33.56 (0.48)</td>
<td>34.56 (0.80)</td>
</tr>
<tr>
<td></td>
<td>1.2</td>
<td>91.50 (0.78)</td>
<td>34.00 (0.77)</td>
<td>34.78 (0.64)</td>
</tr>
<tr>
<td></td>
<td>1.3</td>
<td>88.20 (0.53)</td>
<td>38.72 (0.52)</td>
<td>38.61 (0.80)</td>
</tr>
<tr>
<td></td>
<td>1.4</td>
<td>89.80 (1.10)</td>
<td>39.17 (0.51)</td>
<td>38.22 (0.83)</td>
</tr>
<tr>
<td>Posttest</td>
<td>2.1</td>
<td>89.50 (1.07)</td>
<td>62.44 (0.59)</td>
<td>58.11 (0.46)</td>
</tr>
<tr>
<td></td>
<td>2.2</td>
<td>87.70 (1.40)</td>
<td>62.94 (0.46)</td>
<td>58.67 (0.56)</td>
</tr>
<tr>
<td></td>
<td>2.3</td>
<td>90.60 (0.83)</td>
<td>67.56 (0.47)</td>
<td>60.00 (0.65)</td>
</tr>
<tr>
<td></td>
<td>2.4</td>
<td>89.50 (1.04)</td>
<td>67.56 (0.46)</td>
<td>61.05 (0.56)</td>
</tr>
</tbody>
</table>

*CRPs 1.1, 1.2, 2.1, and 2.2 are chest pain cases; CRPs 1.3, 1.4, 2.3, and 2.4 are stroke cases.

Table 2. Mean CRP scores for each cohort.

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Pretest</th>
<th>Posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMET (1)</td>
<td>36.38 (0.70)</td>
<td>65.12 (0.69)</td>
</tr>
<tr>
<td>COMET (2)</td>
<td>36.58 (0.74)</td>
<td>64.33 (0.57)</td>
</tr>
<tr>
<td>COMET (3)</td>
<td>36.12 (0.76)</td>
<td>65.92 (0.69)</td>
</tr>
<tr>
<td>COMET (all)</td>
<td>36.36 (0.42)</td>
<td>65.12 (0.38)</td>
</tr>
<tr>
<td>Human tutor (1)</td>
<td>36.83 (0.64)</td>
<td>60.96 (0.51)</td>
</tr>
<tr>
<td>Human tutor (2)</td>
<td>37.42 (0.89)</td>
<td>58.63 (0.41)</td>
</tr>
<tr>
<td>Human tutor (3)</td>
<td>35.38 (0.70)</td>
<td>58.79 (0.55)</td>
</tr>
<tr>
<td>Human tutor (all)</td>
<td>36.54 (0.44)</td>
<td>59.46 (0.31)</td>
</tr>
</tbody>
</table>

Acknowledgments

We thank Hugin Expert for letting us use the Hugin Researcher software. Thanks to Thammasat University Medical School for their participation in the data collection and system evaluation.

References


For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.

---

**The Authors**

Siriwan Suebnukarn is an assistant professor at Thammasat University Dental School and an adjunct faculty member of computer science and information management at the Asian Institute of Technology in Bangkok. Her interests lie in intelligent medical training systems, medical and dental informatics, clinical skill training, and problem-based learning. She received her PhD in computer science and information management from the Asian Institute of Technology and her DDS from the Prince of Songkla University. Contact her at the Thammasat Univ. Dental School, Pathumthani, 12121, Thailand; ssiriwan@tu.ac.th.

Peter Haddawy is a professor in the Asian Institute of Technology’s Computer Science and Information Management program, the vice president for the institute’s academic affairs, and a senior scientist at the University of Wisconsin-Milwaukee. His research interests include decision-theoretic problem solving, probabilistic reasoning, preference elicitation, and e-commerce and medical applications. He received his PhD in computer science from the University of Illinois at Urbana-Champaign. Contact him at the Asian Inst. of Technology, School of Eng. and Technology, Pathumthani, 12120, Thailand; haddawy@ait.ac.th.