Fostering Diagnostic Accuracy in a Medical Intelligent Tutoring System

Reza Feyzi-Behnagh, Roger Azevedo, Elizabeth Legowski, Kayse Reitmeyer, Eugene Tseytlin, and Rebecca Crowley

Department of Educational and Counselling Psychology, McGill University, Montreal, Canada
Department of Biomedical Informatics and Pathology, University of Pittsburgh, PA, USA
reza.feyzibehnagh@mail.mcgill.ca; roger.azevedo@mcgill.ca;
{legoex, reitmeyerkl, tseytline, crowleyrs}@upmc.edu

Abstract. Diagnostic classification is an important part of clinical care, which is often the main determinant of treatment and prognosis. Clinicians' under- or over-confidence in their performance on diagnostic tasks can result in diagnostic errors which can lead to delay in appropriate treatment and unnecessary increase in the cost of medical care. This paper presents a version of SlideTutor aiming to reduce pathologists' and dermatopathologists' bias in diagnostic decision-making. This is accomplished by frequently prompting them to make metacognitive judgments of confidence, presenting them with the expert diagnostic solution path for each case, and de-biasing them by making them conscious of their metacognitive biases. This paper describes and summarizes the functionalities of SlideTutor, its cognitive training, tutoring phase, expert feedback, metacognitive intervention, and the open learner model.

1 Introduction and Background

Intelligent tutoring systems (ITSs) are adaptive and personalized instructional systems designed to mimic the well-known advantages of human one-on-one tutoring over other types of instructional methods [e.g., 1]. ITSs are capable of accelerating and enhancing the training of novices by providing adaptive and individualized scaffolding and feedback based on a complex interaction between several modules representing the domain knowledge as well as learner knowledge acquisition and development of expertise. The adaptive scaffolding and feedback in ITSs are targeted at improving student learning and fostering skills, such as making accurate metacognitive judgments [see 2]. In contexts where the teacher has limited time to spend on presenting content, teaching problem solving skills, and providing tailored feedback to individual students, ITSs can prove extremely helpful by providing adaptive individualized instruction to learners, organize content, and point out their errors for as much time and as many iterations as the learner requires [3].

ITSs can prove beneficial in training of highly specialized clinicians, such as pathologists. Training of specialized clinicians is very difficult in traditional training contexts for several reasons, including insufficient exposure to infrequently encoun-

tered cases, and the increased workloads of mentors which limit the time for training the next generation of practitioners and increase the potential for clinical errors among less-experienced practitioners. Training of pathologists typically requires five or more years, which includes both residency training (3-5 years) and advanced fellowship (1-3 years). In the context of training pathologists, ITSs could help alleviate many of the above-mentioned problems by providing a safe environment where residents can practice whenever they have time and as frequently as needed, and receive individualized feedback and guidance without inadvertently harming patients in the process. More specifically, ITSs can scaffold residents' accuracy of diagnoses, thereby alleviating their overconfidence or under-confidence in their performance on diagnostic tasks. Overconfidence would cause the clinician to conclude the diagnosis too quickly, therefore neglecting to fully consider alternative hypotheses and all the evidence in the case, which can result in diagnostic errors [4]. On the other hand, underconfidence might lead them to order unnecessary or inappropriate additional testing and use consultative services, which increases the risk of iatrogonic complications (i.e., complications caused by medical treatment or diagnostic procedures), delays treatment, and unnecessarily increases the costs of medical care [5].

In order to alleviate the problem of under- or overconfidence in residents' diagnostic performance (i.e., poor calibration of judgment and performance), scaffolding needs to be provided to improve the accuracy of their metacognitive judgments (i.e., Feeling of Knowing, FOK) and eliminate any diagnostic bias. FOK is defined as the learner's certainty of his/her actual performance [6]. ITSs can play a significant role in assisting pathologists in making more accurate metacognitive judgments about their diagnostic decision-making and performance, and as a result make more accurate diagnoses.

One of the important methods of scaffolding and improving learners' metacognitive skills and performance is the use of open learner models (OLMs) in ITSs. A student model is an important part of an ITS which observes learner behavior and builds an individualized qualitative representation of her/his cognitive and metacognitive skills and gets updated in real-time during learners' interaction with the ITS [7]. Learner models are usually embedded in the ITS architecture and are not visible to the students, however, several researchers [e.g., 8] have investigated the benefits of allowing learners to access their learner model (OLM). Research has indicated that the mere displaying of visualizations of OLMs in ITS interfaces raises the awareness of the learners, allowing them to reflect on different aspects of their learning and problem solving. Besides all the advantages of using OLMs in interactive ITSs, according to [9], no study has investigated the use of OLMs for displaying metacognitive processes (e.g., metacognitive judgments of correctness of performance). In spite of the great potential and possibilities offered by the use of medical ITSs, few of these systems have been fully developed [e.g., 9] and only a fewer have been empirically evaluated [e.g., 10].

In this paper, we describe an adapted version of SlideTutor, an ITS which scaffolds pathology residents' accuracy of metacognitive judgments using different metacognitive interventions and an OLM for presenting metacognitive accuracy. The paper does not include our evaluation of the effectiveness of the implemented modules.

2 Description of the Medical ITS: SlideTutor

The SlideTutor intelligent tutoring system (http://slidetutor.upmc.edu) was modified for use in this study. The computational methods and the architecture of the original system have been previously published [11]. For the current study, the system uses a modular architecture implemented in the Java programming. SlideTutor provides users with cases to be solved under supervision by the system. Cases incorporate virtual slides, which are gigabyte size image files created from traditional glass slides by concatenating multiple images from a high resolution robotic microscope. Virtual slides are annotated using a custom built editing environment to produce case representations of discrete findings and their locations. A separate Ontology Web Language (OWL) based expert knowledge base consists of a comprehensive set of evidence-diagnosis relationship for the entire domain of study. A reasoning module uses a decision tree approach to construct a dynamic solution graph (DSG), representing the current state of the problem and all acceptable next steps including the best-next-step. As for the interface, participants use a graphical user interface (Fig. 1) to examine and diagnose the cases. Participants can pan and zoom in the virtual slide, locate findings using the mouse, and select from lists of findings and qualifiers, such as size and type, from a tree-like representation. Once findings are specified, they appear as evidence nodes in the diagrammatic reasoning palette (Fig 1). Afterwards, participants assert hypotheses using a separate tree-based menu, which eventually appear as nodes in the diagrammatic reasoning palette. Support links can then be drawn between evidence and hypothesis nodes to specify relationships between the two. Finally, one or more hypotheses may be dragged to the diagnosis window, and selected as the final diagnosis(es) before proceeding to the next case.

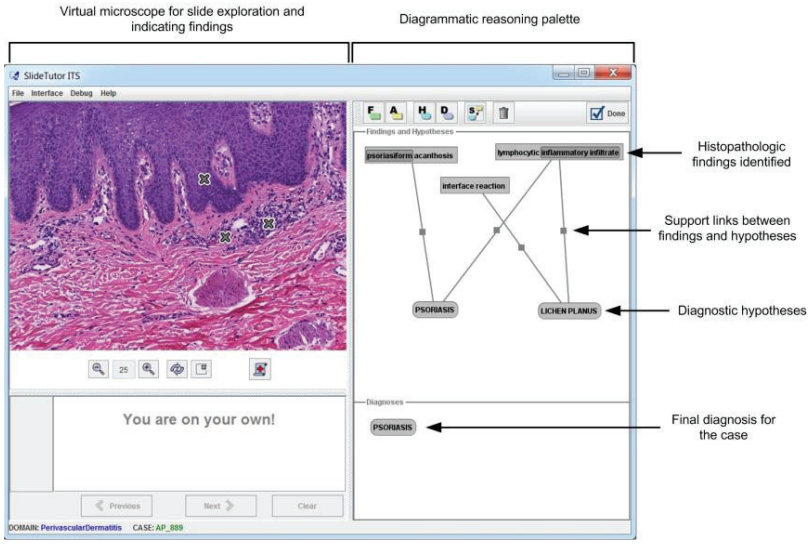


Fig. 1. SlideTutor interface

2.1 The Dynamic Book

An interactive knowledge browser has been developed (called the Dynamic Book) that shows feature-diagnosis relationships as well as glossary information on all features and diagnoses in the selected domain of dermatopathology (i.e., perivascular diseases) (Fig. 2). A description of the domain and the cases is presented in the next section. A total of sixty-two diagnoses and fifty-seven findings are presented in this interface. Six of the diagnoses comprising six patterns were used in the tutoring phase of the study. By clicking on each one of the diagnoses, an image is presented in the interface showing an example of how the disease presents on a patient's skin. A description of the diagnosis was also presented under the image. Additionally, a list of potentially associated findings is presented to the right of the image and diagnosis description. A zoomed-in virtual slide image accompanied each of the findings in the list, where the presentation of the finding is indicated by an arrow. A description of the particular finding together with a list of potentially associated diagnoses is also presented. In order to guide the exploration of participants during the Dynamic Book phase towards important parts of the book, they are provided with a list of tasks to work through which pertained to a mix of patterns they would encounter in the tutoring phase and ones they would not.

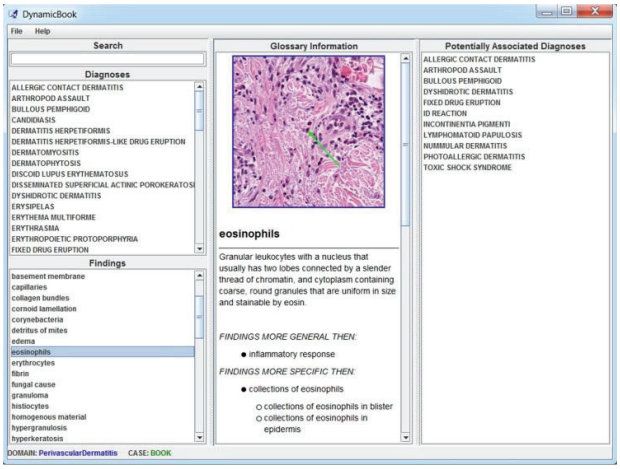


Fig. 2. Dynamic book interface

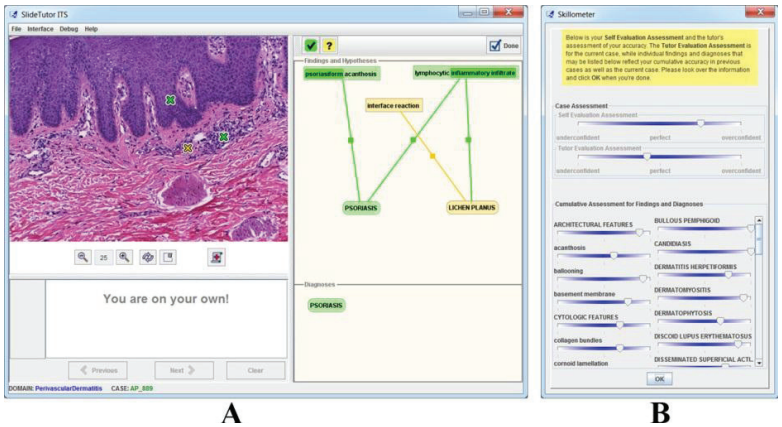
2.2. Pathology Cases

The Perivascular Dermatitis domain was selected for the current SlideTutor study because the domain is well-tested, includes patterns (i.e., a combination of evidence identified in a particular case) with multiple cases, and more cases are available than other domains. Also, Perivascular Dermatitis is a large domain and it is unlikely that participants would have complete knowledge of this diagnostic area. 20 cases were used for the tutoring phase. Cases were obtained from the University of Pittsburgh

Medical Center (UPMC) slide archive and from private slide collections. Diagnoses were checked and confirmed by a dermatopathologist prior to inclusion in the system repository. For each case, a knowledge engineer and an expert dermatopathologist collaborated in defining all present and absent findings, their locations on the slide (case annotation), and relationships among findings and diagnoses (knowledge-base development). Each diagnosis included a set of one or more diseases that matched the histopathologic pattern.

2.3 The Coloring Book and Metacognitive Judgments

For the intervention condition, once participants complete identifying findings, hypotheses, and diagnoses for a case, they progress to an interface called the Coloring Book (Fig. 3A). In this interface, they indicate if they are sure or unsure of the items they identified for the case (i.e., FOK judgments) by clicking on them and coloring them as either green (sure) or yellow (unsure). Next, they are presented with a window with a slider where they indicate how accurate they think their self-assessments in the coloring book were (ranging from underconfident to overconfident). Afterwards, they are presented with correct findings, hypotheses, and diagnoses for the respective case (colored in green) and incorrectly identified items as red. After reflecting on their performance and the feedback from the system, they are presented with a window juxtaposing the sliders for their self-assessment of their FOK judgments and the evaluation of the tutor based on their performance and their FOK judgments (the open learner model: OLM) (Fig 3B). At the bottom of the window, one or more individual findings or diagnoses may be listed, which reflects the participant's cumulative accuracy in previous cases as well as the current case for the particular finding or diagnosis. At the end, they are asked to make another metacognitive judgment and state whether they would feel confident solving similar cases, to which they respond on a 6-point Likert scale ranging from "not confident" to "very confident". This concludes the case, and progresses them to the next case.



A **B**

Fig. 3. Coloring book interface (A) and the OLM (B)

3 Study Timeline

As part of the design of the study and interface of the ITS, the study phases and timeline were determined as follows (Fig. 4). An approximate total time of four hours was allocated as the participant session time. At the beginning and after signing the informed consent form, the participants were administered a test (pre-pre-test) of their prior knowledge of the domain targeted by the current version of SlideTutor (i.e., Perivascular diseases). Next, they spent 30 minutes acquiring cognitive knowledge of the domain while accomplishing a task given to them by the experimenter (Dynamic Book phase). Afterwards, another test of cognitive knowledge of the domain was administered (pre-test). Once the test was completed, they proceeded to the tutor training and tutor use phase (in intervention or control condition) where they solved 20 cases and indicated their confidence in their responses and were shown an OLM (intervention condition), or solved the cases and progressed with no feedback from the system (control condition). At the end, a post-test was administered to gauge their knowledge gains during interactions with the tutor. A detailed description of the ITS, the tests, dynamic book, and the tutoring interventions is presented below.

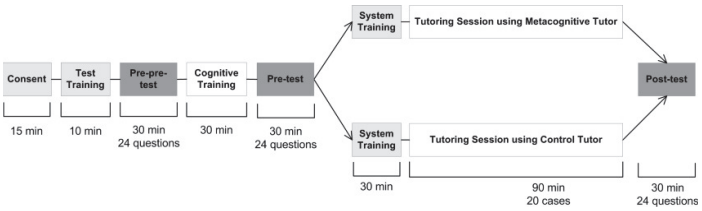


Fig. 4. Study timeline

4 Measures

4.1 Cognitive Measures

In order to measure the prior cognitive knowledge of the domain at the beginning of the tutoring session, cognitive gains after the cognitive learning phase, and the knowledge gains after the tutoring session, three 24-item tests were administered. Three versions of each test were created, and the test order was randomized per session to control for order effects. Each test comprised of 24 questions, and the questions were a mix of tutored and untutored items. Tutored items were about the material that was presented in the cases seen with the tutoring system, while untutored items were about material that was not covered by the tutoring system. Three question types were used in the tests: finding, diagnosis, and differentiate questions. Finding questions consisted of a static microscopic image with an arrow pointing at a feature to be identified. Diagnosis questions consisted of a list of findings, and participants had to provide the diagnosis(es) that match the findings. Differentiate questions consisted of two diagnoses, and participants had to provide a feature that can be used to differenti-

ate the two. After responding to each question, participants were asked to rate if they were sure or unsure of their responses using radio buttons (FOK metacognitive judgment).

4.2 Metacognitive Measures

Feeling of knowing (FOK) metacognitive judgment measures were collected on all test items in the three cognitive knowledge tests and on all findings, hypotheses, and diagnoses identified in cases in the tutoring phase. The FOK measures were collected as binary values: sure vs. unsure. The data from metacognitive ratings on test questions were only used for analyses after the study was completed. However, the metacognitive judgment ratings for items identified in cases in the tutoring phase in the Coloring Book layout (see section 2.3) were used for calculation of a measure of over- or under-confidence called Bias, which was presented to the participant after solving the case and indicated their confidence in the items they identified in the case (in the OLM: see section 2.3). The bias score is calculated by subtracting the relative performance on all items (total correct items divided by all items) from the proportion of items judged as known (total sure items divided by all items) [12]. Figure 5 indicates how bias scores are calculated. Positive bias scores indicate over-confidence and negative scores indicate under-confidence. When performance perfectly matches the rated confidence level, the bias score equals zero. In other words, the bias score indicates the direction and degree of lack of fit between confidence and performance [13]. The bias score for each case was presented to the participant in the form of a slider ranging from under-confident to perfect to over-confident with a cursor indicating the participant's bias score.

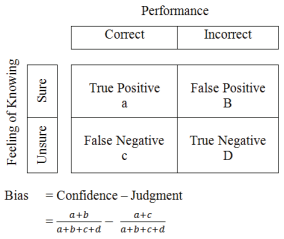


Fig. 5. FOK contingency table and the calculation of bias

5 Conclusion

We described the functionalities of a version of SlideTutor aimed at reducing the metacognitive bias of pathologists and dermatologists while diagnostic decision-making by deploying metacognitive interventions and using an open learner model to aid participants in reflecting on their diagnostic performance. Open learner models have not been used in the previous studies for displaying the metacognitive performance of participants [8], and the current iteration of SlideTutor is novel in this re-

gard. The Dynamic Book interface used for the cognitive learning phase provided participants with an environment to conduct a targeted search and knowledge acquisition (targeted at completing the task assigned by the experimenter). As mentioned above, since the domain chosen for this version of SlideTutor is a very large domain, a cognitive learning phase was deemed necessary in order to provide the opportunity for acquisition of some cognitive knowledge and freely explore the glossary of diagnoses and findings.

Acknowledgment
The authors gratefully acknowledge the support of this research by the National Library of Medicine through grant number 5R01LM007891.

References

1. Koedinger, K., Aleven, V., Roll, I., & Baker, R. (2009). In vivo experiments on whether supporting metacognition in intelligent tutoring systems yields robust learning. In A. Graesser, J. Dunlosky, D. Hacker (Eds.), *Handbook of metacognition in education*. 383-413. Mahwah, NJ: Erlbaum.
2. Azevedo, R., & Hadwin, A. F. (2005). Scaffolding self-regulated learning and metacognition: Implications for the design of computer-based scaffolds. Instructional Science, 33, 367-379.
3. VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197-221.
4. Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic error in internal medicine. Archives of Internal Medicine, 165 (13), 1493-1499.
5. Berner, E. S. & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. The American Journal of Medicine, 121 (5A), S2-S23.
6. Metcalf, J., & Dunlosky, J. (2008). Metamemory. In H. Roediger (Ed.), Cognitive psychology of memory (Vol. 2, pp. 349-362). Oxford: Elsevier.
7. Bull, S. (2004). Supporting Learning with Open Learner Models. Proceedings of 4th Hellenic Conference with International Participation: Information and Communication Technologies in Education, Athens, Greece. Keynote.
8. Bull, S., & Kay, J. (in press). Open learner models as drivers for metacognitive processes. In R. Azevedo & V. Aleven (Eds.). International handbook of metacognition and learning technologies. Amsterdam, The Netherlands: Springer.
9. Azevedo, R., & Lajoie, S. (1998). The cognitive basis for the design of a mammography interpretation tutor. International Journal of Artificial Intelligence in Education, 9, 32-44.
10. El Saadawi, G. M., Tseytlin, E., Legowski, E., Jukic, D., Castine, M., Fine, J., . . . Crowley, R. S. (2008). A natural language intelligent tutoring system for training pathologists: implementation and evaluation. Advances in Health Sciences Education: Theory and Practice, 13, 709-722.
11. Crowley, R. S., & Medvedeva, O. (2006). An intelligent tutoring system for visual classification problem solving. Artificial Intelligence in Medicine, 36 (1), 85-117.
12. Kelemen, W. L., Frost, P. J., & Weaver III, C. A. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. Memory & Cognition, 28(1), 92-107.
13. Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. Metacognition and Learning, 4(1), 33-45.