Poitras, E. G., Lajoie, S. P., Doleck, T., & Jarrell, A. (2016). Subgroup Discovery with User Interaction Data: An Empirically Guided Approach to Improving Intelligent Tutoring Systems. *Educational Technology & Society, 19* (2), 204–214.

Subgroup Discovery with User Interaction Data: An Empirically Guided Approach to Improving Intelligent Tutoring Systems

Eric G. Poitras[1*], Susanne P. Lajoie[2], Tenzin Doleck[2] and Amanda Jarrell[2]
[1]University of Utah, USA // [2]McGill University, Canada // eric.poitras@utah.edu // susanne.lajoie@mcgill.ca // tenzin.doleck@mail.mcgill.ca // amanda.jarrell@mail.mcgill.ca
[*]Corresponding author

(Submitted November 24, 2014; Revised May 12, 2015; Accepted July 31, 2015)

ABSTRACT

Learner modeling, a challenging and complex endeavor, is an important and oft-studied research theme in computer-supported education. From this perspective, Educational Data Mining (EDM) research has focused on modeling and comprehending various dimensions of learning in computer based learning environments (CBLE). Researchers and designers are actively attempting to improve learning systems by incorporating adaptive mechanisms that respond to the varying needs of learners. Recent advances in data mining techniques provide new possibilities and exciting opportunities for developing adaptive systems to better support learners. This study is situated in the context of clinical reasoning in an Intelligent Tutoring System called BioWorld and it aims to examine the relationship between the lab-tests ordered and misconceptions held by learners. Toward this end, we employ an EDM technique called subgroup discovery to unpack the rules that embody the hypothesized link. Examining such links may have implications for identifying the points along learning trajectories where learners should be provided the requisite scaffolding. This study represents our efforts to evaluate and derive empirically based design prescriptions for improving Intelligent Tutoring Systems. Implications for practice and future research directions are also discussed.

Keywords

Subgroup discovery, Educational data mining, Intelligent tutoring systems, Medical education, Clinical reasoning, Lab tests, Misconceptions, Learner modeling

Introduction

Learner modeling is a critical component in the process of adapting instruction to the specific needs of different learners with computerized systems. Adaptive instructional systems may be defined as a systematic process which consists of four steps: (1) capturing information about the learner; (2) analyzing learner interactions through a model of learner characteristics in relation to the domain; (3) selecting the appropriate instructional content and resources; and (4) delivering the content to the learner (Shute & Zapata-Rivera, 2012). The analytical function of the learner model can be further classified in terms of processes conducted at both the macro and micro levels (VanLehn, 2006). At the macro-level, a representation of the path towards competency within the domain is updated for each task with the aim of selecting the next task that is the most appropriate for the learner. At the micro-level, instructional materials such as hints and feedback are delivered to the learner on the basis of a representation that is repeatedly updated over the duration of task performance. Intelligent tutoring systems have been shown to improve upon typical classroom instruction in several disciplines, including mathematics (e.g., Cognitive Tutors; Koedinger & Corbett, 2006), computer science (e.g., Constraint-Based Tutors; Mitrovic, 2003), microbiology (e.g., Narrative-Based Tutors; Rowe, Shores, Mott, & Lester, 2011), as well as physics and computer literacy (e.g., Dialogue-Based Tutors; Graesser, VanLehn, Rosé, Jordan, & Harter, 2001).

The challenge in modeling learning in the context of solving ill-structured problems is that there are multiple paths towards attaining the correct solution (Lajoie, 2003, 2009). These paths should be documented in order to represent common misconceptions or impasses in moving along the trajectory towards competency. In the medical domain, experts have been shown to diagnose patient diseases in different ways, while at the same time, justifying plausible hypotheses on the basis of common evidence items, including patient symptoms and lab test information (Gauthier & Lajoie, 2014; Lajoie, Gauthier, & Lu, 2009). An intelligent tutoring system such as BioWorld can represent these evidence items in terms of a novice-expert overlay system, which compares novice solution paths to those of the experts in order to individualize feedback. A similar modeling technique has been applied in other systems, such as SlideTutor in diagnosing dermatopathology (Feyzi-Behnagh, Azevedo, Legowski, Reitmeyer, Tseytlin, & Crowley, 2014) as well as the MedU virtual patient cases (Berman, Fall, Chessman, Dell, Lang, Leong, Nixon, & Smith, 2011). On the basis of the user interactions that are recorded by BioWorld, the system will highlight areas of

similarities and differences with the expert solution path, allowing novices to self-reflect on their own approach to resolving the problem (Lajoie & Poitras, 2014). Several studies investigating the novice-expert overlay model have been carried out with the aim of capturing linguistic features from written case summaries to tailor feedback content (Poitras, Doleck, & Lajoie, 2014; Lajoie, Poitras, Doleck, & Jarrell, 2015) as well as the impact of goal-orientations and affective reactions towards attention given to feedback (Lajoie, Naismith, Poitras, Hong, Panesso-Cruz, Ranelluci, & Wiseman, 2013).

In this study, we examine the use of subgroup discovery for the induction of rules that characterize the relationship between impasses in problem-solving and lab-tests ordered in BioWorld. Whereas our previous research characterized how experts converged in their paths to solving a problem, the present study captures how novices diverged from an expert solution path. In doing so, we claim that subgroup discovery algorithms are particularly well suited towards describing the multiple paths that characterize problem-solving in ill-structured domains. The findings obtained from this analytical approach have implications for evaluating and deriving empirically-based design guidelines for novice-expert overlay models. In particular, the capabilities of this type of system to provide instruction in becoming aware of common impasses and misconceptions in solving certain types of problems, which is prescribed on the basis of the specific needs of different novices. In the following section, we review the theoretical model that accounts for novices' ability to regulate their efforts in overcoming impasses in problem solving.

The regulation of learning while solving problems in the medical domain

Social cognitive models of self-regulation characterize the process of solving ill-structured problems as a recursive and iterative process, where adjustments to the solution are made on the basis of refining the problem space (Zimmerman & Campillo, 2003). As such, self-regulated problem solvers engage in cycles of forethought, performance, and self-reflection (Zimmerman, 2000). In the forethought phase, novices orient themselves in the problem space, at the same time, formulating a plan to solve the problem. The performance phase is characterized by the novices' efforts to solve the problem by executing the planned steps and monitoring the outcomes. The self-reflection phase involves the novices' evaluations of the overall progress and elaborations about the problem space, resulting in conclusions about the case. In doing so, the problem-solving process is recursive in that the outcomes of prior steps inform the next ones that are taken to solve the problem.

In accordance with the basic phases of self-regulation, novices in the medical domain regulate their own approaches to solving diagnostic problems in accordance with disciplinary-based practices. Self-regulation involves several types of metacognitive activities; namely, orienting, planning, executing, monitoring, evaluating, and elaborating (Meijer, Veenman, & van Hout-Wolters, 2006; Lu & Lajoie, 2008; Lajoie & Lu, 2012). As an example, self-regulated problem solvers have the ability to notice pertinent vital signs, such as the patient heart rate exceeding the normal range, which could potentially be caused by a tumor of the adrenal glands. To test this assumption the plan might entail testing for pheochromocytoma by ordering a lab test to verify serum levels of the catecholamines adrenalin and noradrenalin. This plan is executed, and the lab test was found to be pertinent, as serum levels were elevated, thereby confirming a diagnosis of pheochromocytoma. Self-regulated problem-solvers evaluate their own progress by re-adjusting the plausibility of differential diagnoses. In doing so, a battery of lab tests might be ordered to rule out commonly known alternative diagnoses. The outcomes of these efforts are evaluated and will inform subsequent attempts to solve the problem as the problem-solver becomes progressively more confident in their own solution.

Modeling self-regulation in BioWorld

BioWorld is a computer-based learning environment designed as a cognitive tool (Pea, 1985; Perkins, 1985; Salomon, Perkins, & Globerson, 1991; Derry & Lajoie, 1993; Jonassen & Reeves, 1996; Lajoie, 2000, 2005). Tools embedded in the learning environment aim to support the cognitive and metacognitive activities that mediate performance in diagnosing patient diseases (see Figure 1). In doing so, the system captures learner behaviors that characterize their efforts to regulate several aspects of problem-solving.

Each patient case begins with the case history where novices formulate their differential diagnoses. Once novices select the possible diagnoses that fit with the symptoms, they report their confidence in their primary hypothesis by

using the Belief Meter (% certainty). Novices gather evidence from the case history by highlighting relevant symptoms, the outcome of which is shown in the Evidence Table, which remains visible throughout the problem solving activity. There is a library where novices access additional information about the disease they are investigating. Information in the library represents the typical symptoms and transmission routes of a specific disease, as well as a glossary of medical terminology and diagnostic testing procedures. In order to solve problems, novices order lab tests to confirm or disconfirm specific diagnoses. They do so by ordering tests on the patient chart, where the outcomes of their tests are recorded in the Evidence Table. A consultation tool is present and novices can obtain hints on request, delivered in increasing order of specificity.

A subset of user interactions with interface elements of BioWorld are captured and analyzed by the novice-expert overlay model. These interactions include the evidence items posted to the evidence palette, namely, information regarding patient symptoms and lab-tests ordered in the chart. The user must identify a particular item as pertinent to solving the case by sending it to the evidence palette. The evidence palette thus serves as a monitoring tool, allowing the user to review the evidence items, otherwise referred to as the steps taken to solve the problem. Novices and experts have been found to interact differently with this tool, as monitoring is likely to develop along a trajectory towards becoming more proficient in the domain (Lajoie et al., 2013).

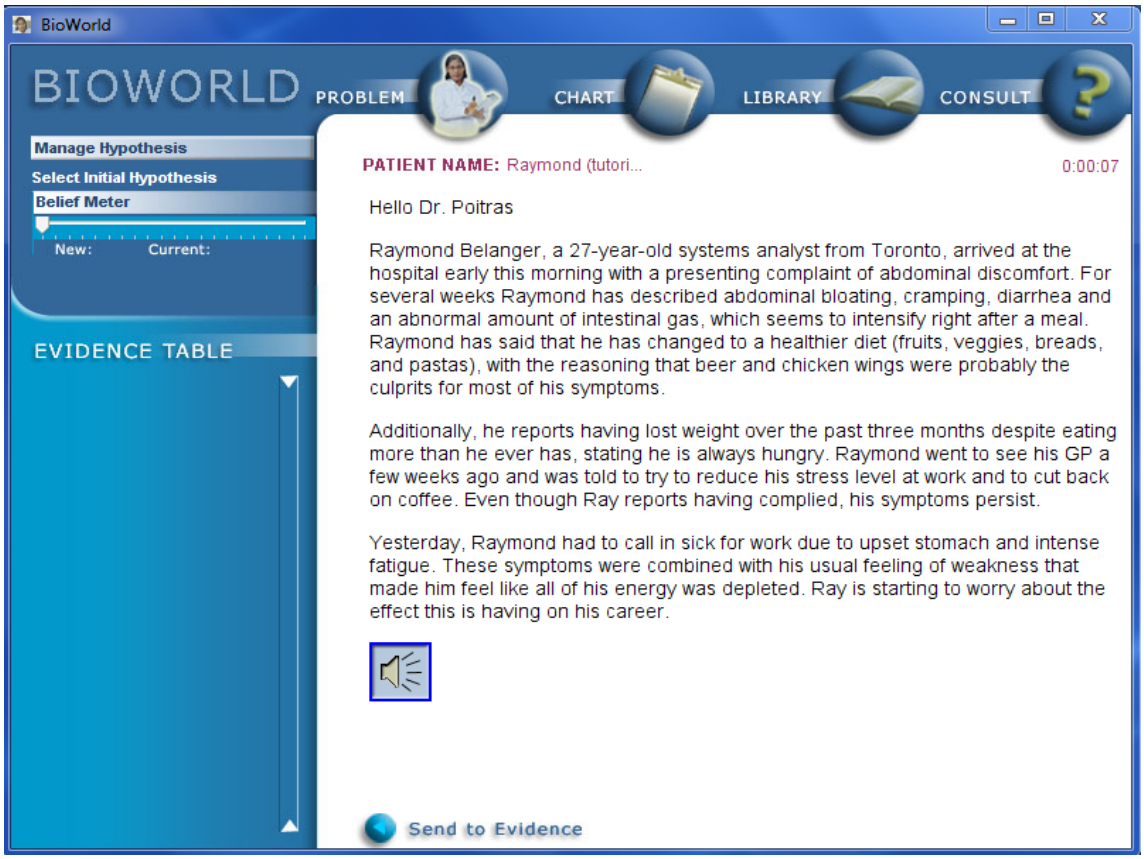

Figure 1. A screenshot of the BioWorld interface

Once novices have submitted their final diagnosis, the novice-expert overlay model individualizes the feedback that is delivered to novices after the submission of their final solution for a particular case. Before novices attend to the feedback, they are required to categorize the evidence items that either confirms, refutes, or that are irrelevant to their final diagnosis. Novices must then prioritize the evidence items in terms of their relative importance in solving the case. In order to foster novices' self-reflection, BioWorld provides them with a case summary and a student report. The case summary consists of a writing activity where novices build a justification for their final diagnosis on the basis of the steps taken to solve the case. The student report refers to the formative feedback that is provided to the novices by the system, where the solution steps of a validated expert solution path is compared to the novices' solution in order to highlight similarities and differences. The expert solution path also provides a case summary,

written by the expert, which outlines in detail the steps that were taken to solve the case, and how each step contributed to formulating the final diagnosis.

Improving the novice-expert overlay model

This study aims to further refine the capabilities of the novice-expert overlay model to individualize instruction by mining learner behaviors to uncover common patterns that are indicative of impasses in problem-solving. Does the application of subgroup discovery to uncover relationships between errors in diagnostic reasoning and the nature of the lab tests that are ordered by novices lead to improvements in the design guidelines of the novice-expert overlay model? We focus on two research questions: (a) What type of diagnostic tests ordered by novices are most often found to precede diagnostic errors during problem-solving?; (b) What are the impacts of these impasses towards diagnostic outcomes as manifested in problem-solving performance and evaluation?

With regard to the first research question, we hypothesize that impasses in problem-solving, as evidenced by ordering non-pertinent lab-tests, are most often found in solving rare and complex cases in BioWorld. As such, the subgroup discovery algorithm is applied to uncover patterns in logged user interactions while solving three cases in BioWorld with varying levels of difficulty. In order to address the second research question, we hypothesize that common impasses encountered while solving a type of disease are indicative of prominent misconceptions about the nature of the disease. We rely on a combination of measures to assess proficiency in diagnostic reasoning as well as novices understanding of the disease after their attempts to solve the problem in order to ascertain the impacts of these impasses in problem-solving.

Methods

Participants

The participants consists of 30 undergraduate students who were compensated $20 for practicing diagnostic reasoning with BioWorld during a 2 hour session. The convenience sample includes 11 men (37%) and 19 women (63%) with an average age of 23 (*SD* = 2.60). Participants were randomly assigned cases to solve, including Amy (Diabetes Mellitus Type 1), Cynthia (Hyperthyroidism), and Susan Taylor (Pheochromocytoma).

Measures

Process measures

BioWorld logs user interactions while interface elements while novices are solving problems. These logs are stored on a MySQL server database in the form of a timestamp, an identifier (i.e., participant and case ID), a space (i.e., BioWorld interface), the user interaction label (e.g., add test), and description (e.g., Thyroid Stimulating Hormone (TSH) Result: 0.2 mU/L). The learning behaviors that were extracted from the database consists of 172 unique lab tests, ordered by the learners using the chart panel in BioWorld. Each variable is represented by its corresponding label (i.e., add test), the name of the procedure (i.e., Abdominal Exam Result), and the value that was obtained (i.e., Normal). These variables are assigned a value that indicates whether the relevant lab test was ordered or not during a particular line of diagnostic reasoning (i.e., true and false, respectively). As such, BioWorld log files were aggregated at the level of changes in the selected or submitted hypothesis for the purposes of determining the occurrence of each learning behavior within each line in diagnostic reasoning. A total of 304 lines of diagnostic reasoning were aggregated by following this procedure.

Product measures

On the basis of the logged user interactions, several performance metrics can be calculated in order to appraise learner performance in solving problems. We distinguish between three types of performance metrics, namely, diagnostic efficacy (e.g., accuracy, count of matches with experts, and percentage of matches with experts),

efficiency (e.g., number of tests ordered and time to solve the case), and affect (e.g., confidence). The target variable that was extracted from the database consists of diagnostic accuracy (i.e., Correct, Incorrect) in order to discover interesting subgroups for the incorrect diagnoses that are indicative of common misconceptions or impasses in solving a case. This target attribute was evenly distributed, with frequencies of 156 correct and 148 incorrect lines of diagnostic reasoning. The accuracy of the diagnosis was determined at the moment when either the learner selected a main hypothesis or the final hypothesis was submitted as the solution for the case. BioWorld logs record changes to the selected or submitted hypotheses, allowing the system to track distinct lines in diagnostic reasoning that are indicative of instances when learners reach an impasse.

Experimental procedure

The data was collected as part of an experiment that examined the antecedent factors to attention allocated towards feedback in BioWorld (see Naismith, 2013). First, participants were asked to complete a demographic questionnaire and the achievement goal questionnaire (i.e., PALS; Midgley, Maehr, Hruda, Anderman, Anderman, Freeman et al., 2000). A training session enabled the participants to learn how to use BioWorld by solving a practice case that was unrelated to the actual cases solved during the study. During the study, the participants performed a think aloud protocol, which involved participants verbalizing their own thought processes while solving a case (see Ericsson & Simon, 1993). At the end of each case, participants filled out the feedback emotions questionnaire (i.e., AEQ; Pekrun, Goetz, Frenzel, Barchfeld, & Perry, 2011). The order of cases to be solved was counterbalanced to mitigate practice effects. The average length of the study was two hours to solve a total of three cases.

Subgroup discovery algorithm

Subgroup discovery (Wrobel, 1997; Klösgen, 2002) is an educational data mining technique that aims to discover interesting rules (i.e., as generalizable as possible and with the most unusual statistical characteristics) with respect to a set of learning behaviors. The technique involves searching for relationships between these different learning variables and a target variable, which are described in terms of individual rules or subgroups. A rule is defined as a conjunction of learning behaviors (i.e., learning behavior-value pair) that is able to account for an unusual statistical distribution in relation to the variable of interest (i.e., target variable). The main difference between the subgroup discovery and the classification task is that subgroup discovery algorithms describe unusual relations between learning behaviors and a certain value of the target variable in a comprehensive manner. By contrast, classification algorithms predict these relationships with an emphasis on precision and interpretability (see Herrera, Carmona, González, & del Jesus, 2011).

Table 1. Quality measures for the subgroup discovery task

Measure	Definition	Formula
Generality		
Positive instances	Count of examples covered by rule and correctly included	N(Antecedent = True & Conclusion = True)
Negative instances	Count of examples covered by rule and incorrectly included	N(Antecedent = True & Conclusion = False)
Size	Count of examples covered by rule	N(Antecedent = True)
Coverage	Percentage of size to the total number of examples	N(Antecedent = True & Conclusion = True)/N(Examples)
Complexity length	Count of variables in antecedent	N(Variables in Antecedent)
Precision		
Precision	Percentage of correct inclusions to the count of examples covered by the rule	N(Antecedent = True & Conclusion = True)/N(Antecedent = True)
Accuracy	Percentage of correct inclusions and exclusions to the total number of examples	[N(Antecedent = True & Conclusion = True)+N(Antecedent = False & Conclusion = False)]/N(Examples)

The subgroup discovery algorithm performs an exhaustive search of the dataset. The dataset consists of the lines of diagnostic reasoning that correspond to a particular case, where the amount of learning behaviors were reduced by calculating the value of the chi-squared statistic with respect to diagnostic accuracy. Learning behaviors whose weights are greater than a null value were selected for inclusion in the subgroup discovery task. The search task was constrained by a number of parameters to identify antecedents that lead to a conclusion of an incorrect diagnosis. The maximum depth of the search was set to 1 in order to limit the number of task iterations, more specifically the number of antecedents of the generated rules. The minimum coverage was set to 5% to ensure that the misconceptions, although rare in occurrence, are still prevalent enough to warrant intervention.

A number of quality measures are used to determine the interpretability of the rules that are generated by the subgroup discovery task, including measures of generality, complexity, and precision. Generality refers to the extent to which examples are covered by the rule. Complexity consists of the simplicity of a particular rule. Precision is defined as the measure of how well a rule identifies and excludes the examples that belong to a subgroup. Table 1 outlines the metrics associated to these quality measures as well as their definitions and formulas.

The main criterion for the evaluation of the quality of subgroups is precision. This criterion is chosen on the basis of the research objectives as it is the most suitable in terms of identifying patterns of learning behaviors that are indicative of misconceptions in diagnostic reasoning. As such, the minimum value for the quality measure of precision was set to 70% in order to determine rules that perform well in terms of including examples to subgroups of misconceptions.

In the following section, we outline the steps involved in the subgroup discovery task. First, we implemented the subgroup discovery algorithm to extract patterns from lines of diagnostic reasoning that relate lab tests ordered to the correctness of the main hypothesis. Second, we validated subgroups of lines of diagnostic reasoning that are characterized by misconceptions using learning outcome measures. Third, a visualization of the lab tests that warranted the main hypotheses was used to facilitate interpretation of the misconceptions. Fourth, decision rules were established to address misconceptions and improve user modeling processes.

Extraction of patterns

Table 2 shows the results obtained by the algorithm with the different parameter values, including the total number of rules obtained, and the values of the quality measures. For the purposes of this analysis, a low number of rules with few attributes is preferred in order to ease the interpretability of the results. Furthermore, the rules are expected to be indicative of common misconceptions when diagnosing diseases. Therefore, the desired result consists of rules that are both precise and accurate in recognizing incorrect lines of diagnostic reasoning on the basis of non-pertinent lab tests ordered by the learners.

Table 2. Patterns extracted from the subgroup discovery algorithm

Antecedent	1	2	3	4	5	6
Incorrect diagnosis for the Cynthia case (i.e., exhibiting signs of Pheochromocytoma)						
Random Blood Glucose Level Result: normal=true	5	1	6	6.1%	83.3%	43.9%
Continual ECG Monitoring Result: Normal=true	9	2	11	11.2%	81.8%	46.9%
Fasting Blood Glucose Level Result: normal=true	5	2	7	7.1%	71.4%	42.9%
ECG with exercise Result: Normal=true	7	3	10	10.2%	70.0%	43.9%

Note. 1 = Positive instances; 2 = Negative instances; 3 = Size; 4 = Coverage; 5 = Precision; 6 = Accuracy.

It is apparent from this table that learners exhibit very few misconceptions while solving the case. A total of 4 rules were extracted to achieve the minimal coverage criteria of 5% of lines of diagnostic reasoning. The range of lines of diagnostic reasoning that were classified as characteristic of a common misconception varied from 6 (6.1%) to a total of 11 (11.2%) lines of diagnostic reasoning. It is noteworthy to mention that rules were only generated in relation to lines of diagnostic reasoning associated to the Cynthia case, which is recognized as the most difficult to solve as Pheochromocytoma is a rare disease. Furthermore, all lab results that are featured as antecedents within the rules are non-pertinent, meaning that the patient vital signs were found to be normal. In sum, misconceptions were only identified in solving more complex cases as opposed to simple ones, as learners attempted but failed to confirm their incorrect diagnoses.

The subgroup discovery rules reveal interesting information about learner misconceptions that can be useful in revising the learner model implemented in BioWorld. As an example, consider the first rule in Table 2. This rule shows that learners who ordered a test to measure Random Blood Glucose levels obtained a value that falls within the normal population range. On the one hand, there are a total of 5 lines of diagnostic reasoning that were characterized by this behavior and that led to an incorrect diagnosis. On the other hand, there was only 1 line of diagnostic reasoning where the learner engaged in this behavior, but still demonstrated a correct understanding of the underlying disease. As such, this particular rule is quite precise, but fairly accurate in characterizing all the lines of diagnostic reasoning included in the dataset (i.e., Precision = 83.3%; Accuracy = 43.9%).

From a medical point of view, higher than normal blood glucose levels may be attributed to rare tumors such as Pheochromocytoma, but also to other causes such as an overactive thyroid gland, pancreatic cancer, pancreatitis, and so on. The most relevant test, however, was to assess the levels of catecholamines through a urine test, as this particular type of tumor usually grows on the adrenal glands. This rule provides new information on how to improve the novice-expert overlay model in BioWorld. It suggests that learners who order this particular test need a decision support system to support them in performing a differential diagnosis. This involves BioWorld supporting the learner in (a) listing the relevant symptoms, (b) the possible diseases that explain the symptom, (c) ruling out possible causes, beginning by removing diagnoses from the list given the non-pertinent lab test result. In this particular case, a learner should rank the likelihood of arrhythmia as less likely given that high blood sugar levels influence heart rhythm. Therefore, arrhythmia would likely be the cause of the symptoms if the lab test showed elevated levels of blood glucose, but this was not the case. The novice-expert overlay model should be programmed to assess these behaviors if the learner orders this particular lab test.

Validation of patterns

Table 3 shows the performance data associated with the outcomes of the learners' efforts to solve the problems in BioWorld. The performance metrics associated with their final solution includes the count and percentage of match with the expert solution, the count of lab test, the level of confidence towards the final solution, as well as the count of correct solutions, incorrect ones, and the ratio of these values. The metrics related to individual lines of diagnostic reasoning during problem-solving consist of the count of correct and incorrect lines of diagnostic reasoning as well as the ratio of these values. Given that the rules generated by the subgroup discovery algorithm are indicative of misconceptions, it is expected that these are associated with lower levels of efficacy, efficiency, and confidence in solving medical problems. In contrast, lines of diagnostic reasoning that are classified by the set of rules as demonstrating no misconceptions should be associated with higher levels of efficacy, efficiency, and confidence.

Table 3. Performance metrics associated to subgroups classified by the subgroup discovery algorithm

Subgroup	1	2	3	4	5	6	7	8	9	10	11
Cynthia case (i.e., exhibiting signs of Pheochromocytoma)											
Misconceptions	18	49%	3.22	14.89	73%	6	12	0.50	7	11	0.64
No Misconceptions	80	55%	3.70	13.65	73%	33	47	0.70	50	30	1.67
Amy case (i.e., exhibiting signs of Diabetes Mellitus (type 1)											
No Misconceptions	105	43%	6.84	10.97	78%	59	46	1.28	79	26	3.04
Susan Taylor case (i.e., exhibiting signs of Hyperthyroidism (Grave's disease)											
No Misconceptions	101	71%	6.22	8.92	86%	58	43	1.35	85	16	5.31

Note. 1 = N; 2 = Average percentage of matches with expert solution; 3 = Average count of matches with expert solution; 4 = Average count of lab tests; 5 = Average level of confidence in case solution; 6 = Count of correct lines of diagnostic reasoning; 7 = Count of incorrect lines of diagnostic reasoning; 8 = Ratio of line correctness; 9 = Count of correct case solutions; 10 = Count of incorrect case solutions; 11 = Ratio of case correctness.

The differences between the subgroups that exhibit either a misconception or lack thereof are highlighted in Table 3 with respect to both the complex and simple cases solved in BioWorld. From this data, it is apparent that the subgroup that exhibited misconceptions when solving the most complex case performed less well than those solving simpler cases. The ratios of correct to incorrect lines of diagnostic reasoning are at least two times higher when solving simple cases. This value doubles to four times higher when appraising the ratios of the correctness of the final case solution for the simple cases. In general, learners who solve simple cases are more confident, order less lab

tests, and have higher counts of matches with the expert solution compared to the subgroup that exhibited misconceptions when solving the complex case.

Although this trend was evident when comparing the subgroup with misconceptions and no misconceptions in relation to solving a complex case, the trend clearly decreases from the levels of performance obtained by the subgroup that solved simple cases. The difference in performance was most notable in ratios correct to incorrect case solutions, where the subgroup with no misconceptions was at least two times more likely to obtain the correct solution. However, the difference between both subgroups is less pronounced in relation to the other performance metrics. There were only slight differences in efficacy, efficiency, and confidence between the subgroups that exhibited misconceptions as opposed to the one that did not while solving the complex case.

Interpretation of patterns

Figure 2 elaborates further on the nature of the misconceptions exhibited in the lines of diagnostic reasoning that were recognized by the subgroup discovery algorithm. In order to further characterize these misconceptions, we examined the main hypothesis selected at the end of each line of diagnostic reasoning that exhibited a misconception. Furthermore, we examined whether the evidence was linked with a particular diagnosis after solving the case, and how the evidence was categorized and prioritized by the learner before BioWorld delivered feedback. The thickness of the lines reflects the frequency count of learners who fell in each category, with thicker lines indicating higher frequencies.

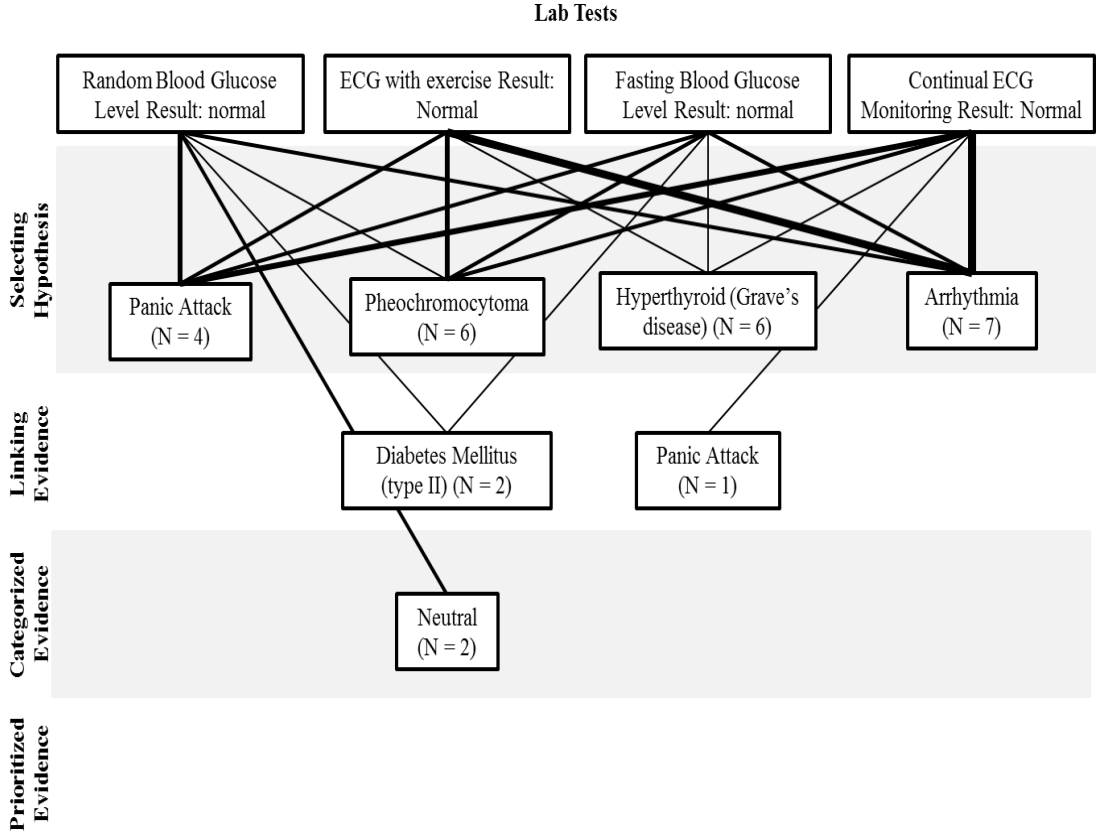

Figure 2. Interpretation of the patterns extracted by the subgroup discovery algorithm

Three broad types of misconceptions emerged from this analysis. When solving a case exhibiting symptoms of Pheochromocytoma, the lines of diagnostic reasoning where a misconception was evident led to incorrect diagnoses of Panic Attack, Hyperthyroid, and Arrhythmia. However, six lines of diagnostic reasoning from a total of 23 within the subgroup that exhibited misconceptions were later followed by the correct diagnosis of Pheochromocytoma. Furthermore, the subgroup rarely engaged in behaviors that were indicative of using the non-pertinent lab test to

warrant erroneous diagnoses, as the frequencies of linking, categorizing, and prioritizing these evidences were relatively low. This finding suggests that although these misconceptions are commonly encountered while solving the complex case, they are not necessarily constant or immutable. Lines of diagnostic reasoning unfold as a dynamic process as the learners' confidence in multiple diagnoses shift on the basis of available symptoms, lab tests, and library information. Given the fact that misconceptions are characterized as a transient condition, their early detection by BioWorld is the first step to delivering remedial instruction that will support learners in becoming more competent.

Conclusion

Learner modeling in computer based learning environments has become a central challenge for instructional technology designers and researchers. Learners encounter obstacles and hold misconceptions that can hinder learning. Therefore, it is imperative to support learners in their individual learning trajectories by embedding scaffolding and feedback through computer based cognitive tools. A chief obstacle to the design of adaptive learning environments is to first identify individual differences and to respond to misconceptions. The redesign of functionalities and response mechanisms in learning environments will substantially contribute to this objective.

The purpose of this study was to develop an understanding of misconceptions in medical problem solving by exploring pertinent and non-pertinent laboratory tests in BioWorld, with the goal of adapting instruction and providing requisite scaffolding. To address this purpose we used subgroup discovery to generate rules, which provided evidence about the relationship between misconceptions in clinical reasoning and laboratory tests ordered in BioWorld. The results suggested that subgroup discovery algorithms are particularly well suited towards ascertaining the difficulties that characterize problem-solving in ill-structured domains such as medical problem solving. In particular, the analyses revealed the specific antecedents of incorrect diagnoses, with regards to the laboratory tests ordered. Performance metrics were associated with correct performance as well as specific misconceptions for each of the cases. The results from this study have implications for evaluating and deriving empirically-based design guidelines with regards to tailoring instruction and scaffolding to individual learners. For example, these findings will contribute towards improving the metacognitive support provided to help learners become aware of misconceptions in solving medical problems.

It is worth considering some limitations of this investigation. First, there are many factors that can contribute to the decision of ordering certain lab-tests, of which we only considered a few. In future studies, we will systematically analyze learner rationales for ordering certain lab-tests by examining think out loud protocols. Second, we constrained our analysis to a small number of rules with few attributes to ease the interpretability of the results. We will conduct an expanded analysis to generate a complete set of rules with all the attributes to establish a holistic picture of misconceptions during medical problem solving. Third, our analyses were limited to three medical problems, which challenges the generalizability of our findings. Extracting patterns from a larger set of clinical cases will lead to a better understanding of misconceptions and more precise generation of rules across diverse medical problems.

The use of subgroup discovery is a promising method to reveal and understand the relationship between misconceptions in clinical reasoning and types of laboratory tests ordered. There is still much to be gained from examining the misconceptions that learners encounter and the ways to help support learners overcome these difficulties. Future work will consist of investigating alternative approaches for delineating the relationship between obstacles in clinical reasoning and lab-tests ordered. More importantly, we hope to incorporate the lessons gleaned from this study to make the novice-expert overlay model in BioWorld more robust to provide adaptive scaffolding for learners engaging in clinical reasoning. This work is an important step towards redesigning computer based learning environments to support learners on their individual learning trajectories.

Acknowledgements

We gratefully acknowledge the contributions of Laura Naismith and Maedeh Kazemitabar in collecting data for this project. This research is funded by the Social Sciences and Humanities Research Council of Canada.

References

Berman, N. B., Fall, L. H., Chessman, A. W., Dell, M. R., Lang, V. R., Leong, S. L., Nixon, L. J., & Smith, S. (2011). A Collaborative model for developing and maintaining virtual patients for medical education. *Med Teach, 33*(4), 319-324.

Derry, S. J., & Lajoie, S. P. (Eds.) (1993). *Computers as cognitive tools*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Feyzi-Behnagh, R., Azevedo, R., Legowski, E., Reitmeyer, K., Tseytlin, E., & Crowley, R. S. (2014). Metacognitive scaffolds improve self-judgements of accuracy in a medical intelligent tutoring system. *Instructional Science, 42*(2), 159-181.

Gauthier, G., & Lajoie, S. P. (2014). Do expert clinical teachers have a shared understanding of what constitutes a competent reasoning performance in case-based teaching? *Instructional Science, 42*(4), 579-594. doi:10.1007/s11251-013-9290-5

Graesser, A. C., VanLehn, K., Rosé, C. P., Jordan, P. W., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI magazine, 22*(4), 39.

Herrera, F., Carmona, C. J., González, P., & del Jesus, M. J. (2011). An overview on subgroup discovery: Foundations and applications. *Knowledge and information systems, 29*(3), 495-525. doi:10.1007/s10115-010-0356-2

Jonassen, D. H., & Reeves, T. C. (1996). Learning with technology: Using computers as cognitive tools. In D. H. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 693-719). New York, NY: Macmillan.

Klösgen, W. (2002). Types and forms of knowledge (patterns): Subgroup patterns. In W. Klösgen & J. Zytkow (Eds.), *Handbook of Data Mining and Knowledge Discovery* (pp. 47-51). New York, NY: Oxford University Press.

Koedinger, K. R., & Corbett, A. (2006). Cognitive tutors: Technology bringing learning sciences to the classroom. In K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences* (pp. 61-78). New York, NY: Cambridge University Press.

Lajoie, S. P. (2000). *Computers as cognitive tools: Vol.2. No more walls: Theory change, paradigm shifts and their influence on the use of computers for instructional purposes*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lajoie, S. P. (2003). Transitions and trajectories for studies of expertise. *Educational Researcher, 32*(8), 21–25.

Lajoie, S. P. (2005). Cognitive tools for the mind: The Promises of technology: Cognitive amplifiers or bionic prosthetics? In R. J. Sternberg & D. Preiss (Eds.), *Intelligence and Technology: Impact of Tools on the Nature and Development of Human Skills* (pp. 87-102). Mahwah, NJ: Erlbaum.

Lajoie, S. P. (2009). Developing professional expertise with a cognitive apprenticeship model: Examples from Avionics and Medicine. In K. A. Ericsson (Ed.), *Development of professional expertise: Toward measurement of expert performance and design of optimal learning environments* (pp. 61-83). New York, NY: Cambridge University Press.

Lajoie, S. P., & Poitras, E. (2014). Macro and micro-strategies in for metacognition and co-regulation in the medical tutoring domain. In R. Sottilare, A. Graesser, X. Hu, & B. Goldberg (Eds.), *Design Recommendations for Adaptive Intelligent Tutoring Systems: Adaptive Instructional Management (Volume 2)* (pp. 151-168). Orlando, FL: U.S. Army Research Laboratory.

Lajoie, S. P., Gauthier, G., & Lu, J. (2009). Convergence of data sources in the analysis of complex learning environments. *Research and Practice in Technology Enhanced Learning, 4*(3), 195–219.

Lajoie, S. P., & Lu, J. (2012). Supporting collaboration with technology: Does shared cognition lead to co-regulation in medicine? *Metacognition and Learning*, (7), 45-62.

Lajoie, S., Naismith, L., Poitras, E., Hong, Y. J., Cruz-Panesso, I., Ranellucci, J., Mamane, S., & Wiseman, J. (2013). Technology rich tools to support self-regulated learning and performance in medicine. In R. Azevedo & V. Aleven (Eds.). *International Handbook of Metacognition and Learning Technologies* (pp. 229-242). New York, NY: Springer.

Lajoie, S. P., Poitras, E. G., Doleck, T., & Jarrell, A. (2015). Modeling metacognitive activities in medical problem-solving with BioWorld. In A. Peña-Ayala (Ed.), *Metacognition: Fundamentals, Applications, and Trends. A Profile of the Current State-Of-The-Art* (pp. 323-343). doi:10.1007/978-3-319-11062-2_13

Lu, J., & Lajoie, S. P. (2008). Supporting medical decision making with argumentation tools. *Contemporary Educational Psychology, 33*, 425-442.

Meijer, J., Veenman, M. V. J., & van Hout-Wolters, B. H. A. M. (2006). Metacognitive activities in text-studying and problem-solving: Development of a taxonomy. *Educational Research & Evaluation, 12*(3), 209-237.

Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., Gheen, R., Middleton, M. J., Nelson, J., Roeser, R., & Urdan, T. (2000). *Manual for the Patterns of Adaptive Learning Scales (PALS)*. Ann Arbor, MI: University of Michigan.

Mitrovic, A. (2003). An Intelligent SQL tutor on the web. *International Journal of Artificial Intelligence in Education*, *13*(2), 173-197.

Naismith, L. (2013). *Examining motivational and emotional influences on medical students' attention to feedback in a technology-rich environment for learning clinical reasoning* (Unpublished doctoral dissertation). McGill University, Canada.

Pea, R. D. (1985). Beyond amplification: Using the computer to reorganize mental functioning. *Educational Psychologist, 20*(4), 167-182.

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance. The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology, 36*(1), 36–48.

Perkins, D. N. (1985). The Fingertip effect: How information-processing technology shapes thinking. *Educational Researcher*, *14*(7), 11-17.

Poitras, E., Doleck, T., & Lajoie, S. (2014). Mining case summaries in BioWorld. In *Proceedings of the 9th International Conference on Computer Science & Education (ICCSE 2014)* (pp. 6-9). doi:10.1109/ICCSE.2014.6926421

Rowe, J. P., Shores, L. R., Mott, B. W., & Lester, J. C. (2011). Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education, 21*(1), 115-133.

Salomon, G., Perkins, D. N., & Globerson, T. (1991). Partners in cognition: Extending human intelligence with intelligent technologies. *Educational researcher*, *20*(3), 2-9.

Shute, V. J., & Zapata-Rivera, D. (2012). Adaptive educational systems. In P. Durlach & A. Lesgold (Eds.), *Adaptive technologies for training and education* (pp. 7-27). New York, NY: Cambridge University Press.

VanLehn, K. (2006). The Behavior of tutoring systems. *International Journal of Artificial Intelligence in Education, 16*(3), 227-265.

Wrobel, S. (1997). An Algorithm for multi-relational discovery of subgroups. In *Proceedings of the first European symposium on principles of data mining and knowledge discovery (PKDD-97)* (pp. 78–87). New York, NY: Springer.

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). San Diego, CA: Academic Press.

Zimmerman, B. J., & Campillo, M. (2003). Motivating self-regulated problem solvers. In J. E., Davidson & R. Sternberg (Eds.), *The Nature of Problem Solving* (pp. 233-262). New York, NY: Cambridge University Press.