# Using Learning Analytics to Identify Medical Student Misconceptions in an Online Virtual Patient Environment

Eric G. Poitras
*University of Utah*

Laura M. Naismith
*University Health Network*

Tenzin Doleck, Susanne P. Lajoie
*McGill University*

## Abstract

This study aimed to identify misconceptions in medical student knowledge by mining user interactions in the MedU online learning environment. Data from 13000 attempts at a single virtual patient case were extracted from the MedU MySQL database. A subgroup discovery method was applied to identify patterns in learner-generated annotations and responses to multiple-choice items on the diagnosis and management of acute myocardial infarction (i.e., heart attack). First, the algorithm generated rules where single terms from the learner annotations were used to predict incorrect answers to the multiple-choice items. Second, the possible combinations of terms and their relevant synonyms were used to determine whether their inclusion led to better rates of prediction. The second step was found to significantly increase prediction precision and weighted relative accuracy, uncovering four misconceptions at a rate greater than 70%. These findings serve to inform the design of an adaptive system that tailors the delivery of formative feedback to promote better learning outcomes in the domain of clinical reasoning.

## Introduction

Many students enter medical school with misconceptions that are resistant to change. A study of first- and second-year medical students showed that half of first-year students held one or more misconceptions prior to attending a course on the cardiovascular system, with this number only decreasing slightly in their second year of study (Ahopelto, Mikkalä-Erdmann, Olkinuora, & Kääpä, 2011). Students who held such misconceptions performed poorly on a related clinical reasoning task, supporting the fundamental role of knowledge in developing clinical reasoning expertise (Norman, 2005). In the absence of targeted instruction and detailed feedback, even more advanced medical students can struggle to debug their own misconceptions and restructure their knowledge appropriately (Boshuizen, van de Wiel, & Schmidt, 2012). Amongst practicing physicians, cognitive errors related to faulty prior knowledge or synthesis have been shown to contribute significantly to diagnostic error (Graber, Franklin, & Gordon, 2005). Furthermore, Graber et al. (2005) found that early errors influenced future errors, highlighting the need to identify and address misconceptions as early as possible in the learning process.

Misconceptions that lead to diagnostic errors can be challenging to examine in the clinical environment (Norman & Eva, 2010). In addition to providing valuable opportunities for students to practice and get feedback on their clinical reasoning skills, online learning environments can support instructors and researchers in tracking and logging user interactions to understand learning processes. Learning analytic techniques can be used to leverage the affordances of the large volumes of data generated by learners in these environments. The application of learning analytics to individualize instruction is referred to as learner modeling (Baker & Siemens, 2014; Ellaway, Pusic, Gallbraith, & Cameron, 2014; Ferguson, 2012; Kay, Reimann, Diebold, & Kummerfeld, 2013). Learner models allow instructional systems to capture and analyze user interactions in order to select and deliver the most suitable instructional content (Shute & Zapata-Rivera, 2012) and personalize the delivery of feedback (Feyzi Behnagh et al., 2014; Lajoie et al., 2013) to enhance learning outcomes.

In our previous research, we investigated the utility of a learning analytic technique known as subgroup discovery (Wrobel, 1997; Klösgen, 2002), a data-driven approach for generating rules that describe subsets of a population that are both sufficiently large and statistically unusual. We used this approach to classify learner interactions in another computer learning environment (Lajoie, 2009) where students learn to reason about virtual patient cases by formulating diagnoses and ordering laboratory tests. By examining the laboratory tests selected by learners in the context of diagnosing a virtual patient, we were able to generate rules that were suggestive of relationships between specific laboratory tests ordered and misconceptions in clinical reasoning (Poitras, Lajoie, Doleck, & Jarrell, 2016). However, the applicability of this approach for discovering knowledge from unstructured text-based data has yet to be determined. In the current study, we investigated the suitability of subgroup discovery methods for characterizing relationships between learner-generated text annotations and their subsequent responses to multiple-choice questions in the MedU online learning environment (Fall et al., 2005). Our aim was to examine the specific linguistic features of annotations associated with incorrect answers to identify common misconceptions, and to use such data to inform the provision of dynamic feedback in the context of problem-solving.

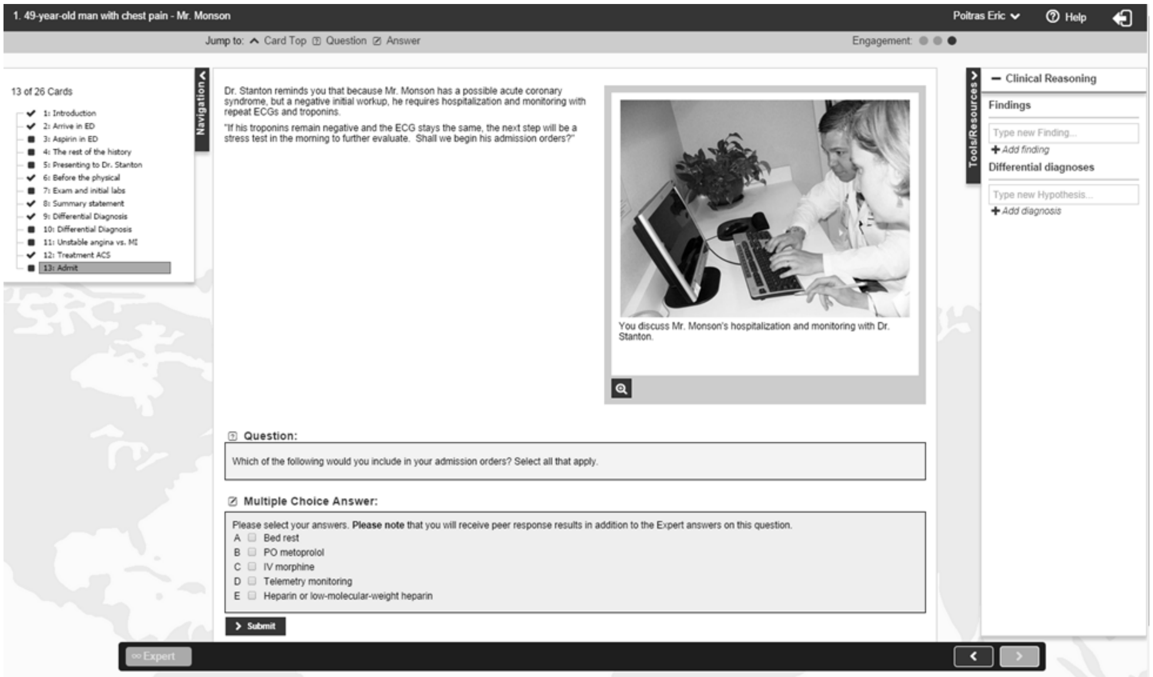
## Detecting Misconceptions in Online Learning

Learning outcomes in online learning environments can be measured in terms of the degree of alignment between a learner model and an expert model of performance. Misconceptions and missing conceptions (Van Lehn, 1988) represent different types of discrepancies between these two models: missing conceptions refer to knowledge that the expert model contains that the student model does not, while misconceptions refer to knowledge that the student model contains and the expert model does not. Once detected, these discrepancies can be addressed in different ways. In the *model tracing approach* (Anderson, Boyle, Corbett, & Lewis, 1990), the learner model is continually compared against the expert model. If a discrepancy is identified, the learner is immediately and directly prompted to perform particular actions. In contrast, the use of *novice-expert overlay models* (Lajoie, Poitras, Doleck, & Jarrell, 2015) as a feedback mechanism puts the onus on learners to determine their next actions. The novice-expert overlay approach appears to be particularly effective in ill-structured domains where there may be multiple paths to obtaining the correct solution (Lajoie, 2003). Feedback that addresses learner misconceptions can be provided when erroneous solution paths are identified. Examples of the use of novice-expert overlay models to promote expertise in clinical reasoning include the Diagnostic Pathfinder (Danielson et al., 2007), the NUDOV system (Wahlgren, Edelbring, Fors, Hindbeck, & Stahle, 2006), and BioWorld (Lajoie et al., 2013, 2015). In this study, we extended the prior analytics methods we used with BioWorld to analyze learner actions in the more widely-used MedU online learning environment.

## Virtual Patient Cases in MedU

MedU (www.med-u.org) has a current user base of 30,000 learners across North America and contains sets of virtual patient cases targeted at various specialties including pediatrics, internal medicine, family medicine, and radiology (Fall et al., 2005). A virtual patient case in MedU consists of a series of interactive HTML screens, or “cards”. Each card presents the learner with new information, including patient symptoms, current vital signs, electronic medical records, and consultation notes from other physicians. The sequential arrangement of the cards is intended to simulate how the condition of the patient evolves over time and in response to actions taken by the learner. The progression of patient symptoms is also made evident through explicit discussions of hospitalization, patient management plans, and treatment outcomes. The MedU environment embeds a number of tools to support learners in formulating their own differential diagnoses and treatment plans (see Figure 1). The *Navigation* sidebar allows learners to view the full set of cards in a case and navigate back to a previously viewed card. The *Tools/Resources* sidebar contains three free-text input fields where learners can record key findings, differential diagnoses and make annotations pertaining to the case.

MedU does not contain explicit learner or expert models. Instead, learner performance is analyzed through their responses to multiple-choice questions embedded throughout the case. These questions may address underlying knowledge and/or prompt learners to choose amongst a set of options for performing an action. The learner responses are highlighted as correct or incorrect and the appropriate actions that need to be taken are reviewed (e.g., “You should call a cardiology consult immediately”) and justified (e.g., “Urgency is critical as you want to prevent further myocardial damage…”). At the same time, the expert palette provides hints that are helpful in performing the correct action, as in a list of criteria to diagnose a left bundle branch block on an electrocardiogram (e.g., “The heart rhythm must be supraventricular in origin…”). Recent studies have sought to expand the nature of assessment in MedU. For example, Smith et al. (2016) developed a rubric for human tutors to assess students’ written case summary statements in MedU. We contend that learner annotations represent an important and as yet untapped source of data for detecting and addressing learner misconceptions.



Dr. Stanton reminds you that because Mr. Monson has a possible acute coronary syndrome, but a negative initial workup, he requires hospitalization and monitoring with repeat ECGs and troponins.

"If his troponins remain negative and the ECG stays the same, the next step will be a stress test in the morning to further evaluate. Shall we begin his admission orders?"

You discuss Mr. Monson's hospitalization and monitoring with Dr. Stanton.

Question:

Which of the following would you include in your admission orders? Select all that apply.

Multiple Choice Answer:

Please select your answers. **Please note** that you will receive peer response results in addition to the Expert answers on this question.

A □ Bed rest
B □ PO metoprolol
C □ IV morphine
D □ Telemetry monitoring
E □ Heparin or low-molecular-weight heparin

*Figure 1*. A screenshot of the MedU user interface.

### Research Objectives

The aim of this study was to detect learner misconceptions in the context of solving MedU virtual patient cases. Specifically, we applied a subgroup discovery method to search for common patterns in the learner-annotation data that were predictive of incorrect answers to multiple-choice questions. In doing so, our objective was to gain insights into the nature of misconceptions that mediate diagnostic performance and should therefore be targeted for corrective feedback. Our specific research questions included:

(1) Which linguistic features of learner annotations are associated with errors?
(2) Does the use of multiple terms increase prediction quality compared to single terms?
(3) Can the onset of misconceptions be detected?

## Method

### Materials

**Dataset.** We developed our analysis using the first case of the Structured Internal Medicine Patient Learning Experience (SIMPLE) set of MedU cases. In this case, Mr. Monson, a 49-year-old man, presents to the emergency department with an acute episode of chest pain, nausea and shortness of breath (Figure 1). The case walks the learner through the diagnosis and management of acute myocardial infarction (i.e., heart attack). This case consisted of 24 cards, 8 of which contained multiple-choice questions. With approval from the institutional review board at McGill University, we extracted anonymized student performance data for this case from the MedU MySQL database. The number of attempts for each multiple-choice question ranged from 12947 to 13262 ($M$ = 13067, $SD$ = 90). As the multiple-choice questions were designed such that multiple responses were required (i.e., “check all that apply”), we evaluated each item selection separately, for a total of 45 true/false items. For example, the first multiple-choice question asks students to select immediate treatment actions and provides four possible responses: (a) aspirin, (b) electrocardiogram (ECG), (c) sublingual nitroglycerin, and (d) a lab draw for cardiac troponins. These options were labelled as true/false items 1.1, 1.2, 1.3, and 1.4 for analysis purposes.

Sixty-seven percent of learners correctly identified acute myocardial infarction as a possible differential diagnosis (8786 correct, 4360 incorrect). This difficulty level suggested that some misconceptions were likely to be present.

**Data scoring and coding**. An example of a learner annotation made while using the Tools/Resources toolbar is shown below:

> *associated symptoms include nausea, mild epigastric tenderness dyspnea on exertion; relieved after 5 minutes of rest patient has history of tobacco use patient has family history of heart disease CXR normal no rubs, gallops, bruits RRR somatoform disorder pancreatitis GERD musculoskeletal - pulled muscle Angina not sx: no sweating, nausea, vomiting, radiating pain MI not in pain currently ischemic arterial disease - CAD pain in center of chest*

We extracted linguistic features of the annotations through a series of text pre-processing steps. The annotations were first isolated into terms by tokenizing the full string using non-characters such as spaces. For example, in the annotation above, the words “associated,” “symptoms,” “include,” and “nausea” were each represented as separate terms. The terms were then transformed to lower case and filtered on the basis of length to exclude terms less than 2 characters. Any terms corresponding to stop-words such as *the*, *is*, *at*, *which*, and *on* were also excluded. A vector was created for each annotation using the complete set of terms extracted. Each cell in the vector contained a binary value representing term occurrence (i.e., true = 1) or non-occurrence (i.e., false = 0). The complete session-by-term vector was then pruned using a threshold of 10% in order to exclude terms with rare occurrences and reduce the dimensionality of the dataset.

**Data analysis**. Subgroup discovery involves an exhaustive search for relationships between learner behaviors, which consist of a set of predictor variables and a target variable of interest. In this study, the target variable in the dataset was the user selection for each of the 45 true/false items (i.e., binary value of 1 if selected or 0 if not selected). The subgroup discovery task was set to generate rules that account for incorrect responses to each item (i.e., true or false, depending on the item selected) on the basis of the linguistic features of learners' annotations to the Tools/Resources toolbar. For example, in the expert response to the first multiple-choice question, items a (1.1), b (1.2), and d (1.4) were selected. As our objective was to model incorrect responses, the target variables for the subgroup discovery algorithm were set to 0 (i.e., not selected) for items 1.1, 1.2, and 1.4, and 1 (i.e., selected) for item 1.3.

We applied the subgroup discovery algorithm to the MedU dataset using a two-step supervised approach. To facilitate analysis, the search task was constrained by a number of parameters. For example, the minimum coverage of a rule was set to 1% to ensure that errors in responding to the true/false items were prevalent enough to warrant intervention. In step one, the algorithm performed an exhaustive search across the dataset with the maximum depth of search set to 1, such that each rule contained a single-term *antecedent* from a learner annotation (e.g., nausea=true), that was associated with a *conclusion* (e.g., item 1.1=incorrect). We optimized for precision of rule detection. The number of task iterations was limited to select the set of 10 rules that was most precise in detecting errors for each true/false item. In step 2, the 10 antecedents from step 1 as well as relevant synonyms selected from an exhaustive word list generated from the text mining method were analyzed with the maximum depth of the search increased to 10. In doing so, the possible combination of terms and synonyms, up to a maximum of 10, were tested in order to determine whether a more precise solution could be obtained. All data analyses were performed in RapidMiner Studio, version 6.4 (https://rapidminer.com).

There are number of quality measures used in subgroup discovery (Herrera, Carmona, Gonzalez, & Jose del Jesus, 2011) to quantify the statistical novelty of a subgroup or rule (Konijn, Duivesteijn, Meeng & Knobbe, 2014). Duivesteijn and Arno (2011) note that selecting the right quality measure is often a difficult task; this choice is either driven by familiarity with the measures or based on default choice. For the present study, we calculated four common quality measures for each rule: Accuracy, Coverage, Precision, and Weighted Relative Accuracy (WRAcc; Lavrač, Flach, & Zupan, 2000). Accuracy was a measure of the proportion of instances where the rule made the correct prediction: (true positives + true negatives) / (all instances), while Precision represented the proportion of correct rule predictions amongst all cases where the antecedent term was found: (true positives) / (all positives). Coverage was calculated by dividing the number of annotations where the antecedent term was found by the size of the complete data set: (all positives) / (all instances). The WRAcc metric balanced both coverage and accuracy to assess the novelty of a subgroup relative to the entire population (Lavrač et al., 2000). WRAcc was calculated by subtracting the product of the probabilities that either the antecedent or conclusion were true from the probability that both the antecedent and conclusion were true: P (antecedent=true & conclusion=true) - [P(antecedent=true) X P(conclusion=true)]. WRAcc values can range from -.25 to .25, with values near 0 representing little novelty compared to the entire data set.

For each of the 45 true/false items, we calculated the average for each quality measure across the 10 rules generated for each step of the analysis. Paired *t*-tests were used to assess whether there were statistically significant improvements to any of the quality measures from step 1 to step 2 of the analysis. We calculated item difficulty by dividing the number of incorrect answers from the total number of item responses. The item difficulty value is thus independent from the quality measures used to appraise the rules generated through the subgroup discovery mining algorithm, and serves to corroborate the findings obtained from the analysis. In an attempt to facilitate early detection of misconceptions, we also mined the HTML content of the cards by extracting terms and matching them with a list of the antecedent terms generated by the subgroup discovery method.

# Results

## Summary Findings

After pruning to exclude rare terms, a total of 821 terms from the set of learner annotations were analyzed. Table 1 shows the descriptive statistics for each performance metric across all 45 true/false items. From step 1 to step 2 of the subgroup discovery algorithm, there was a significant increase in the precision of detecting errors, *t*(44) = -10.73, *p* < .0005 and in the weighted relative accuracy, *t*(44) = -2.61, *p* = .01, but no significant change to either accuracy, *t*(44) = -1.62, *p* = .11, or coverage, *t*(44) = 1.42, *p* = .16.

Table 1
*Means and Standard Deviations for Subgroup Discovery Algorithm Quality Metrics (N = 45)*

| Metric | Step 1 | | Step 2 | |
|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* |
| Accuracy | 72.4% | 19.3% | 72.6% | 19.4% |
| Precision | 32.4% | 21.7% | 35.8% | 22.5% |
| Coverage | 1.9% | .70% | 1.6% | 1.2% |
| Weighted Relative Accuracy | .11% | .10% | .15% | .14% |

Figure 2 and 3 shows the average precision and weighted relative accuracy values for each true/false item. Four items (2.3, 2.9, 3.4, and 3.5) were associated with average precision values greater than 70%. All of these items also had relatively high difficulty scores, ranging from 62% to 76%. In the following section, we elaborate on the most difficult item (i.e., 3.4) to illustrate the rule antecedents generated by the subgroup discovery task and the calculations of the different quality metrics.

Using Learning Analytics to Identify Medical Student Misconceptions in an Online Virtual Patient Environment

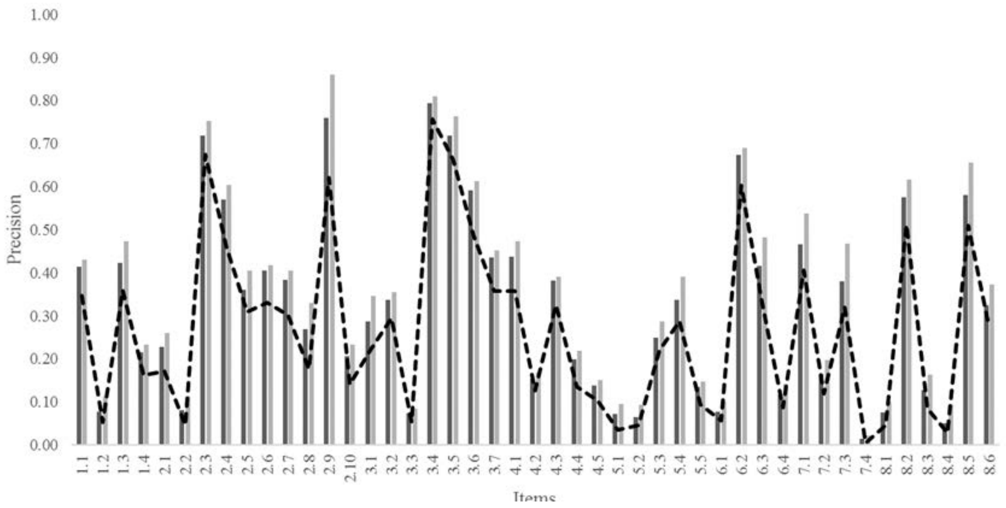


*Figure 2*. Average precision values for step 1 and step 2 across all 45 items for the Mr. Monson case.

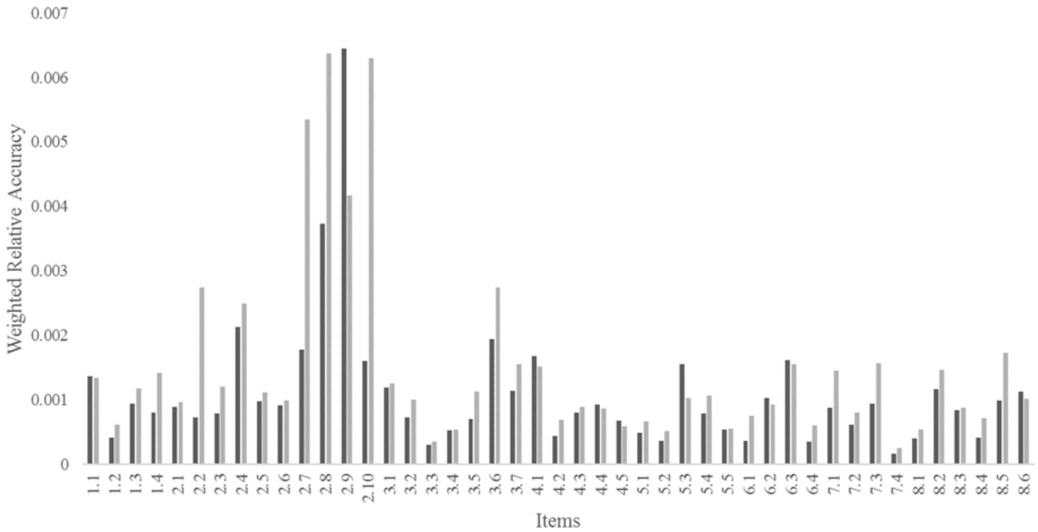


*Figure 3*. Weighted relative accuracy values for step 1 and step 2 across all 45 items for the Mr. Monson case.

## Detailed Item Analysis

We detected a potential misconception related to the third multiple-choice question, which asked the learner to select the appropriate treatment for the management of a patient with ongoing chest pain due to unstable angina. Item 3.4 was especially difficult for learners. This item referred to the anticoagulant drug Heparin, which was selected (i.e. “true”) in the expert response as being pertinent to the treatment plan. Of the 13073 learner responses to item 3.4 in the database, 9903 (76%) were classified as incorrect. The antecedents identified from the learner annotations in step 1 of the analysis included the terms: “segment,” “neg,” “reflux,” “central,” “felt,” “carrying,” “worse,” “yrs.,” “induced,” and “dad.” For example, of the 134 instances where the term “segment” was found in a learner's annotations, 108 responses to item 3.4 were also found to be incorrect, resulting in a detection precision of 80.6% (i.e., 108/134). As shown in Table 2, precision values increased in step 2 when multiple terms were used as antecedents.

In contrast to the high precision value, the accuracy of the term “segment” in detecting an incorrect response to item 3.4 was only 24.9% (i.e., [108 + 3144]/13073). This means that while mentioning “segment” was likely to be associated with an incorrect response to item 3.4, not mentioning “segment” was not necessarily indicative of a correct response to this item. As such, the mention of the antecedent term was a sufficient factor for inferring learner misconceptions, but not a necessary one. The coverage of this rule across the entire data set was 1.0% (i.e., 134/13073), meaning that misconceptions were typically rare amongst learners. Accordingly, the weighted relative accuracy was also low at .05% (i.e., 108/13073-[134/13073 X 9903/13073]).

A subset of the antecedents listed in Table 2 were found in the HTML content of the cards for this case. Item 3.4 appeared on the 12th card. The term “induced” was only mentioned on the first card (i.e., “Cocaine-induced chest pain”) as a potential cause of chest pain related to the cardiovascular system. However, these terms figured prominently in the subsequent card where the patient complained about his symptoms:

> *…So, while I was pulling carpet out and* ***carrying*** *it up the stairs, I noticed some discomfort in my chest. It was not really severe; it* ***felt*** *like some pressure. In retrospect, I think it might have been heartburn because I* ***felt*** *a little nauseated too…*

Table 2

*Identified Rules from the Subgroup Discovery Task for Item 3.4*

| Antecedents | Tru | Fal | Cov | WRAcc | Pre |
|---|---|---|---|---|---|
| Step 1 | | | | | |
| segment=true | 108 | 26 | 0.01 | 4.97E-04 | 0.81 |
| neg=true | 149 | 37 | 0.01 | 6.20E-04 | 0.80 |
| reflux=true | 115 | 29 | 0.01 | 4.53E-04 | 0.80 |
| central=true | 110 | 28 | 0.01 | 4.18E-04 | 0.80 |
| felt=true | 156 | 41 | 0.02 | 5.18E-04 | 0.79 |
| carrying=true | 109 | 29 | 0.01 | 3.41E-04 | 0.79 |
| worse=true | 124 | 33 | 0.01 | 3.88E-04 | 0.79 |
| yrs=true | 138 | 37 | 0.01 | 4.16E-04 | 0.79 |
| induced=true | 141 | 38 | 0.01 | 4.13E-04 | 0.79 |
| dad=true | 421 | 114 | 0.04 | 0.001203 | 0.79 |
| Step 2 | | | | | |
| negative=false AND reflux=true AND yr=false | 109 | 25 | 0.01 | 5.73E-04 | 0.81 |
| neg=false AND negative=false AND reflux=true AND yr=false | 108 | 25 | 0.01 | 5.55E-04 | 0.81 |
| negative=false AND reflux=true | 112 | 26 | 0.01 | 5.71E-04 | 0.81 |
| neg=false AND negative=false AND reflux=true | 111 | 26 | 0.01 | 5.52E-04 | 0.81 |
| negative=false AND reflux=true AND segment=false AND yr=false | 106 | 25 | 0.01 | 5.18E-04 | 0.81 |
| negative=false AND reflux=true AND worse=false AND yr=false | 106 | 25 | 0.01 | 5.18E-04 | 0.81 |
| neg=false AND negative=false AND reflux=true AND worse=false AND yr=false | 106 | 25 | 0.01 | 5.18E-04 | 0.81 |
| negative=false AND reflux=true AND segment=false | 109 | 26 | 0.01 | 5.15E-04 | 0.81 |
| negative=false AND reflux=true AND worse=false | 109 | 26 | 0.01 | 5.15E-04 | 0.81 |
| neg=false AND negative=false AND reflux=true AND worse=false | 109 | 26 | 0.01 | 5.15E-04 | 0.81 |

Notes: True Positives (Tru), False Positives (Fal), Coverage (Cov), Weighted Relative Accuracy (WRAcc), Precision (Pre)

Furthermore, these terms were also mentioned in the context of feedback provided to learners:

> *…Most patients presenting with chest pain should have an ECG done immediately to look for ST* ***segment*** *abnormalities that indicate myocardial injury. ST* ***segment*** *elevations are present in a STEMI; in a NSTEMI, ST* ***segment*** *depressions may occur or the ST* ***segments*** *may be normal…*

Table 3 shows the terms that could be traced back to the HTML content of each card.

Table 3
*Count of Item 3.4 Antecedent Term Mentions in the HTML Cards for the MedU Case*

| Antecedents | Case Card Number | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| segment=true | | 5 | | | | | | | | | 1 | |
| neg=true | | | | | | | | | | | | |
| reflux=true | | | | | | | | | | | | |
| central=true | | | | | | | | | | | | |
| felt=true | | 3 | | | | | | | | | | |
| carrying=true | | 1 | | 1 | | | | | | | | |
| worse=true | | | | 1 | | | | | | | | |
| yrs=true | | | | | | | | | | | | |
| induced=true | 1 | | | | | | | | | | | |
| dad=true | | 1 | | | | | | | | | | |

## Discussion

This study examined the application of learning analytic techniques to learner-generated annotation data in MedU, an online virtual patient environment. Based on data from a single case completed by over 13000 actual users, we were able to use the subgroup discovery technique to automatically generate rules to associate learner annotations with incorrect responses to 45 true/false items. From step 1 (single antecedent terms) to step 2 (multiple antecedent terms), we noted a significant increase in the precision and weighted relative accuracy of error detection, but no significant impact on either coverage or accuracy.

Responding to Ellaway et al.'s (2010) caution on the risk of high false positives when using large data sets, we purposely selected the parameters of our analysis to generate rules with a high precision of error detection (i.e., a low false positive rate). The high correlation between item difficulty and precision values obtained validated the choice of precision as the optimizing metric. We contend that the rules associated with precision values greater than 70% are likely to represent misconceptions, where the learner model contains knowledge that the expert model does not. For example, the results from step 1 suggest that an annotation of "reflux" was four times more likely to be associated with an incorrect answer to item 3.4 than a correct answer. All 10 rules generated in step 2 for this item also included the term "reflux" (see Table 2). Taken together, these results suggest that students may attribute Mr. Monson's symptoms to gastroesophageal reflux disease, a common condition that mimics myocardial infarction (Mayo Clinic, 2015). Conversely, the low overall accuracy of the rules can be attributed to a high false negative rate where the rule fails to detect incorrect answers. The low accuracy rate suggests that subgroup discovery may be less effective in detecting missing conceptions, where the expert model contains knowledge that the learner model does not. The low coverage and weighted relative accuracy values suggest that the detected misconceptions were relatively uncommon, and therefore unlikely to be detected by other means.

While automated techniques can successfully identify misconceptions, detecting their onset was considerably more difficult. By investigating the occurrence of the antecedent terms in the HTML content of the case, we were able to hypothesize about the nature of learner misconceptions. For example, the majority of the antecedent terms for item 3.4 appeared early in the HTML content of the case (see Table

3). Learners who used these terms in their annotations may have had difficulty updating their knowledge as the case evolved. Upon further investigation, we found that gastroesophageal reflux disease was a correct differential diagnosis for this case and was identified as such by 69% of learners (9088 correct, 4058 incorrect). Instead of a knowledge misconception, the inclusion of “reflux” as an annotation may reflect a tendency towards the cognitive error of premature closure, where other possibilities are not considered once an initial diagnosis is made (Graber et al., 2005). A complementary approach to searching the HTML content for the antecedent terms would be to investigate the context of the annotations in which the terms were mentioned. Both of these sources of information would allow domain experts to interpret the nature of the misconception in more depth.

Once identified, different approaches to providing feedback may be needed for different types of misconceptions. Model tracing approaches (e.g., Anderson et al., 1990) may be more effective for resolving misconceptions related to underlying knowledge, while novice-expert overlay approaches are likely to be more effective for supporting learners in self-regulating their cognitive processes while problem-solving (Lajoie et al., 2013).

## Significance

Learning analytic techniques such as subgroup discovery provide an unprecedented opportunity to use data from real learners in authentic learning situations to better understand learning processes. This study illustrates how automated techniques can be used to detect learner misconceptions, formulate theory-based hypotheses about their sources, and inform the provision of personalized feedback to promote better learning outcomes.

## Acknowledgements

This research was supported by a grant from Med-U. The content is solely the responsibility of the authors and does not necessarily represent the official view of Med-U.

## References

Ahopelto, I., Mikkilä-Erdmann, M., Olkinuora, E., & Kääpä, P. (2011). A follow-up study of medical students' biomedical understanding and clinical reasoning concerning the cardiovascular system. *Advances in Health Sciences Education, 16,* 655-668.

Anderson, J. R., Boyle, C. F., Corbett, A. T., & Lewis, M. W. (1990). Cognitive modeling and intelligent tutoring, *Artificial Intelligence, 42*, 7-49.

Baker, R., & Siemens, G. (2014). Educational data mining and learning analytics. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (2nd ed., pp. 253-272). Cambridge, UK: Cambridge University Press.

Boshuizen, H. P. A., van de Wiel, M. W. J., & Schmidt, H. G. (2012). What and how advanced medical students learn from reasoning through multiple cases. *Instructional Science*, *40*, 755-768.

Danielson, J. A., Mills, E. M., Vermeer, P. J., Preast, V. A., Young, K. M., Christopher, M. M., et al. (2007). Characteristics of a cognitive tool that helps students learn diagnostic problem solving. *Educational Technology, Research and Development, 55*(5), 499-520.

Duivesteijn, W., & Arno, A. (2011). Exploiting false discoveries–Statistical validation of patterns and quality measures in subgroup discovery. In *IEEE 11th International Conference on Data Mining (ICDM)* (pp. 151-160). Vancouver, BC: IEEE.

Ellaway, R. H., Pusic, M. V., Galbraith, R. M., & Cameron, T. (2014). Developing the role of big data and analytics in health professions education. *Medical Teacher, 36*(3), 216-222.

Fall, L. H., Berman, N. B., Smith, S., White, C. B., Woodhead, J. C., & Olson, A. L. (2005). Multi-institutional development and utilization of a computer-assisted learning program for the pediatrics clerkship: the CLIPP Project. *Academic Medicine, 80*(9), 847-855.

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning, 4*(5-6), 304-317.

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

Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic error in internal medicine. *Archives of Internal Medicine, 165*, 1493-1499.

Herrera, F., Carmona, C. J., Gonzalez, P., & Jose del Jesus, M. (2011). An overview on subgroup discovery: Foundations and applications. *Knowledge and information systems, 29*(3), 495-525.

Kay, J., Reimann, P., Diebold, E., & Kummerfeld, B. (2013). MOOCs: So many learners, so much potential. *IEEE Intelligent Systems, 28*(3), 70-77.

Klösgen, W. (2002). Subgroup discovery. In W. Klösgen and J. Zytkow (Eds.), *Handbook of data mining and knowledge discovery*. New York: Oxford University Press.

Konijn, R., Duivesteijn, W., Meeng, M., & Knobbe, A. (2014). Cost-based quality measures in subgroup discovery. *Journal of Intelligent Information Systems*, 1-19.

Lajoie, S. (2003). Extending the scaffolding metaphor. *Instructional Science, 33*(5), 541-557.

Lajoie, S. (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: Cambridge University Press.

Lajoie, S., Naismith, L., Poitras, E., Hong, Y., Panesso-Cruz, I., Ranelluci, J., & 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*. Amsterdam: The Netherlands: Springer.

Lajoie, S., Poitras, E., Doleck, T., & Jarrell, A. (2015). Modeling metacognitive activities in medical problem-solving with BioWorld. In Peña-Ayala (Ed.), *Metacognition: Fundaments, Applications, and Trends.* Springer Series: Intelligent Systems Reference Library.

Lavrač, N., Flach, P., & Zupan, B. (2000). Rule evaluation measures: A unifying view. In *Inductive Logic Programming, Lecture Notes in Computer Science, vol. 1634* (pp. 174-185). Berlin: Springer.

Mayo Clinic (2015). Diseases and Conditions: Heartburn. http://www.mayoclinic.org/diseases-conditions/heartburn/in-depth/heartburn-gerd/art-20046483 (accessed July 17, 2015).

Norman, G. (2005). Research in clinical reasoning: Past history and current trends. *Medical Education, 39*(4), 418-427.

Norman, G. R., & Eva, K. W. (2010). Diagnostic error and clinical reasoning. *Medical Education*, *44*(1), 94-100.

Poitras, E., Lajoie, S., 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.

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

Smith, S., Kogan, J. R., Berman, N. B., Dell, M. S., Brock, D. M., & Robins, L. S. (2016). The development and preliminary validation of a rubric to assess medical students' written summary statements in virtual patient cases. *Academic Medicine, 91*(1), 94-100.

Van Lehn, K. (1988). Toward a theory of impasse-driven learning. In H. Mandl & A. Lesgold (Eds.), Learning issues for intelligent tutoring systems (pp. 19-4 1). New York: Springer-Verlag.

Wahlgren, C.-F., Edelbring, S., Fors, U., Hindbeck, H., & Stahle, M. (2006). Evaluation of an interactive case simulation system in dermatology and venereology for medical students. *BMC Medical Education*, *6*, 40.

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* (pp. 78-87). New York: Springer.