Thomspon, G., Morrison, R.G., Holyoak, K.J. & Clark, T.K. (2006). Evaluation of an online analogical patient simulation program. *Proceedings of the Nineteenth IEEE International Symposium on Computer-Based Medical Systems,* (pp. 623-628). Los Alamitos, CA: IEEE Computer Society.

Evaluation of an Online Analogical Patient Simulation Program

Gregory A. Thompson, MD
Medantic Technology,
greg@medantic.com

Robert G. Morrison, PhD
Xunesis
robertmorrison@xunesis.org

Keith J. Holyoak, PhD
UCLA Dept. of Psychology
holyoak@lifesci.ucla.edu

Terry K. Clark, MD
Medantic Technology,
terry@medantic.com

Abstract

Medulator™, *a commercial Web-based, variable response, patient simulation application, was modified to test the effect of case sequencing, explicit case comparison, and user-generated case summaries on overall user performance. METHODS: Senior medical students completed analogous sets of virtual patient cases in different sequences, and their case performance was tracked. A follow-up user satisfaction survey was conducted. RESULTS: A significant effect of case sequencing on analogy transfer was seen only with respect to correct treatment scores ($p = .009$). Explicit case comparison had no reliable effect on performance. However, diagnostic accuracy increased ($p \leq .002$) while treatment attempts decreased ($p = .05$) when subjects were prompted to write case summaries. Student satisfaction with the patient simulation program was high. CONCLUSION: Manipulating case sequences and supporting explicit case comparison yielded mixed results. However, using case summaries as a tool for reflection and proxy for self-explanation led to significant improvement in students' performance.*

1. Introduction

Analogical (case-based) reasoning is ubiquitous in real-world medical diagnosis, and yet most technology-enhanced medical training systems fail to provide new information and training in a manner consistent with the way professionals need to later access learned information. ***Medulator*** (Medantic Technology, Salt Lake City, UT, USA) provides an ecologically valid alternative to these systems and also provides the opportunity to optimize learning through application of principles of transfer from the analogical reasoning literature. Numerous laboratory studies have suggested conditions that may facilitate transfer via analogy [see 1–3 for reviews]; however, few of these methods have been evaluated in complex learning environments such as the domain of medicine. In the present study we use ***Medulator*** to evaluate two potential methods for the optimization of learning.

Analogical reasoning involves comparison of structured information (i.e. the pattern of relationships) between two cases and allows the reasoner to make inferences about one case (the *target*) based on prior knowledge of another case (the *source*). For example, the diagnoses for two patients may be said to be analogous if they have similar patterns of symptoms and diagnostic test results. However, the objects in the source and target of an analogy can also be similar at a surface level (e.g., two patients may be the same gender, race, age or have a similar occupation). These non-diagnostic surface characteristics can frequently be quite salient and can distract reasoners from a full appreciation of the structural similarities between two cases. Thus, transfer to novel cases will be promoted if the learner is led to focus on structural similarities. There is some debate in the experimental literature as to whether surface similarities that correlate with structural similarities may aid in initial learning. On one hand, the salience of surface similarities may facilitate initial detection of the less salient structural similarities, at least for young children [4]; however, the presence of these surface similarities may in some circumstances lead the learner to overlook the diagnostic structural characteristics [5]. In this study we investigate these alternatives by varying the order of cases with respect to surface and structural similarity.

A second factor that has been shown to affect analogical transfer and learning under certain circumstances is explicit case comparison during study [6–8]. For example, Gentner, Lowenstein and Thompson [8] demonstrated that business students were more likely to recall analogically relevant source cases when they were required to explicitly compare

cases during study. However, the effectiveness of this strategy may be domain-specific, both because of the way knowledge in a domain is structured, and also because different types of learners may implicitly use analogy as a standard learning mechanism. For instance, medical and legal professionals who work in a domain that is dominated by case-based reasoning may not be as sensitive to explicit comparison enhancement as business students who work in a domain that is not as structured with respect to cases. To evaluate the effect of explicit comparison during study, we have modified the ***Medulator*** Final Assessment section to include an explicit Analogy Transfer Evaluation (ATE). The ATE requires learners to compare and contrast the current case to previous known cases (at least one of which is a true structural analog of the current case).

An integral part of ATE is the case summary component, which students use as a self-reminder of previously solved cases' germane features when comparing and contrasting to an unknown case. Literature also shows that using self-explanation in problem solving tasks improves performance [9–12]. In the context of ATE, case summary serves as a self-explanation proxy. Thus, a separate arm of this study examined the effect of user-generated case summaries on user performance, independent of ATE.

2. Research Objectives

1. Determine whether case ordering that manipulates the relative surface and structural similarity between adjacent cases affects learning as measured by ***Medulator*** performance metrics.
2. Determine whether explicit comparison as implemented through the ATE can enhance learning as measured by ***Medulator*** performance metrics.
3. Determine effect of case summaries on learning as measured by ***Medulator*** performance metrics.
4. Determine user satisfaction with ***Medulator*** and perceived effect of ATE on diagnostic process.

3. Methods

We used ***Medulator*** to study the effect of case sequencing, explicit case comparison, and writing case summaries on diagnostic and treatment performance using cases that systematically varied structure (i.e., diagnosis determinants such as full symptom constellation, physical examination findings, diagnostic test results, response to therapy, etc.) and surface characteristics (i.e., salient, non-diagnostic information such as patient age, gender, occupation, chief complaint, and presenting symptoms). Participants were senior medical students.

Each participant worked through 11 physician-authored ***Medulator*** virtual patient cases. Diagnoses were from one of three groups: (1) four analogous bioterrorism cases of primary lower respiratory infections (anthrax, pneumonic plague, Q fever, and tularemic pneumonia), (2) four analogous cardiology cases of congestive heart failure (hypertensive CHF, idiopathic dilated cardiomyopathy with CHF, acute MI with CHF, and infective endocarditis with CHF), and (3) three non-analogous distracter cases. Cases were structurally analogous within their own diagnostic category but superficially similar within and/or across diagnostic categories.

The same 11 Cases were presented in one of two different orderings. In the "easy" ordering, cases that had similar structural and surface characteristics were presented earlier in the sequence, while in "hard" ordering early cases shared only structural characteristics and not surface characteristics. As a result both groups of participants saw identical cases and most importantly, the critical test cases (#9 and #11) were identical between groups.

Ninety six (96) senior medical students who had never used ***Medulator*** self-enrolled over the Web and were paid $150 for their participation which took 5.0 hours on average. Participants were randomly assigned to one of six groups (see Table 1).

Table 1. Randomized study group assignment

Group	Case Ordering	ATE	Case Summary
1	Hard	Yes	Yes
2	Hard	No	Yes
3	Hard	No	No
4	Easy	Yes	Yes
5	Easy	No	Yes
6	Easy	No	No

Participants completed cases in a defined sequence (Easy or Hard). For Groups 1 and 4, explicit case comparison was invoked via Analogy Transfer Evaluation (ATE) on 2nd, 3rd, and 4th (test case) analogs of each analogous case set; in these two groups, students were encouraged to write a case summary to serve as a reminder of the cases' salient features. All other groups received no instructions for comparing previous cases but proceeded directly to the Final Assessment section (final diagnosis and treatment selections). However, Groups 2 and 5 also wrote case summaries (without ATE) for each case while Groups 3 and 6 did not. Subjects wrote case

summaries by typing into a free-text space and submitting for later retrieval.

The measured dependent variables were:

1. Number of treatment attempts required to achieve a positive patient outcome
2. Number of correct/incorrect diagnoses chosen
3. Number of correct/incorrect treatments chosen
4. Number/total costs of diagnostic tests chosen
5. Case time (keyboard time)

Participants were initially allowed 2 weeks to complete the study. However, in order to achieve the goal of at least 64 completions, some subjects were granted up to 3 one-week extensions. An honor system was published stating that subjects would work independently and with no external assistance.

3.1 ATE Methodology

The ATE condition consisted of 2 parts:

1. First, once students completed selecting their final diagnoses and final treatments, they were presented with instructions to rate the degree to which previously completed cases were structurally similar to the current case. Upon submitting their ratings, students were given feedback as to which case(s) were the closest analogs (as determined by the case authors).
2. Next, students were asked to select the categories in which the case analogs were most similar, then most different. Eight structural categories were offered for comparison and contrast, including symptom constellation, pertinent diagnostic tests, effective treatments, etc. Students were then asked to justify their responses in free text. Upon submission of this page, an expert opinion of the analogies was given, which students were expected to use to mentally index the current analog for future reference.
3. Finally, the correct diagnoses and treatments for the case were revealed with feedback on the student's selections.

3.2 Satisfaction Survey Methods

Following completion of the 11 cases, subjects were asked to complete a simple user satisfaction survey online. All subjects were asked 4 core questions with Likert scale responses. ATE subjects were asked an additional 2 questions specifically related to their ATE experience.

4. Results

Of the 96 subjects who originally self-enrolled, 72 subjects completed all 11 cases (33 in ATE group, 39 controls). Only data from subjects who completed all 11 cases were analyzed. Outlying data were discarded using a three standard deviation cut.

Case Ordering. To measure the effect of case order (Easy vs. Hard) on performance we analyzed test cases 9 (a bioterrorism case) and case 11 (a cardiology case). To control for the effect of case summaries groups 3 and 6 were eliminated from analysis; thus all participants in this analysis wrote case summaries for every case. Half of the participants in this analysis compared cases using the ATE instructions (groups 1 and 4) while half just performed case summaries (groups 2 and 5). We performed a 2 x 2 between-subjects ANOVA to investigate the effect of case sequencing and explicit comparison (ATE) as well as their possible interaction. Though main effects of case order and ATE were not statistically reliable, a reliable cross-over interaction between case order and ATE was seen ($p = .009$) on the correct treatment score (see Figure 1 below). If case type (bioterriorism vs. cardiology) is included in the analysis, there is a trend ($p = .13$) towards a 3-way interaction, which appears to be driven more by the cardiology case than by the bioterrorism case.

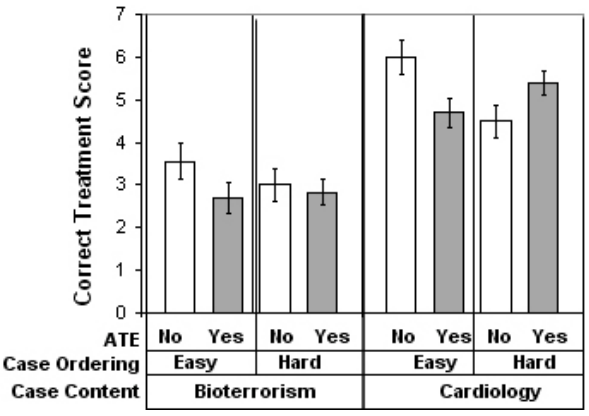

Figure 1. Three-way interaction of case type with case ordering and Analogy Transfer Evaluation (ATE) procedure. Error bars represent ± 1 standard error.

Explicit Case Comparison (ATE). To evaluate the effectiveness of explicit comparison between cases as implemented using the ATE, we only included cases in which the participants in the ATE group were given ATE instructions (thus the first case in each diagnostic category as well as the distracter cases were eliminated). Groups who did not write case summaries (Groups 3 and 6) were not included in this analysis. We performed a mixed 2 x 2 ANOVA which factored

diagnostic category (i.e., cardiology vs. bioterrorism cases) as a within-subject factor and explicit comparison (ATE vs. no-ATE) as a between subject variable. Participants spent more on tests (i.e., test costs) on bioterrorism cases ($p = .03$) and there was a trend ($p = .05$) toward a reduction in test costs in the ATE condition. Likewise, diagnosis, treatment, and quiz performance were all harder with bioterrorism cases ($p < .001$) but there was no reliable effect of ATE. Treatment attempts (i.e. number of measured treatment attempts users required to achieve a positive patient outcome) were higher with ATE ($p = .013$). ATE produced no significant effect on study and solution time (case time).

Case Summaries. To measure the effect of case summary, independent of ATE, we looked at all cardiology and bioterrorism cases. Groups 2 and 4 were eliminated from analysis because they did ATE in addition to case summaries. We performed a mixed 2 x 2 ANOVA which factored diagnostic category (i.e., cardiology vs. bioterrorism cases) as a within-subject factor and whether or not participants were required to write case summaries as a between subject variable. Diagnostic accuracy reliably increased ($p \leq .002$) while treatment attempts decreased ($p = .05$) when participants were prompted to write case summaries. A trend suggested that writing case summaries improved diagnosis of bioterrorism cases more than cardiology cases ($p = .06$) and treatment attempts mainly decreased with case summary writing for cardiology cases ($p = .04$). Writing case summaries had no reliable effect on test costs, or treatment score.

4.1 User Satisfaction Survey Results

Seventy one (71) subjects completed the satisfaction survey, 33 of which were ATE subjects.

Sixty five percent (65%) of subjects responded that they have used similar patient simulation tools only rarely (once per year or less) or never. Overall, 93% rated ***Medulator*** 4 (very good) or 5 (excellent), while 73.2% rated 4 (highly) or 5 (completely) for applicability of ***Medulator*** to their training. User comments indicated that applicability was reduced due to the concentration of bioterrorism cases rather than more "common" medical cases.

Ninety seven percent (97%) of ATE participants responded that ***Medulator*** was moderately or very effective at helping them to think analogically and 75.8% felt that using analogical reasoning was moderately or very effective in helping them to solve cases.

5. Discussion

In this study we attempted to apply several principles from the analogy basic research literature to enhance medical learning using ***Medulator***, an online, interactive, multimedia patient simulator. Specifically, we investigated whether using explicit comparison of cases through Analogy Transfer Evaluation (ATE) would increase students' ability to identify relevant diagnostic and therapeutic principles, improving their clinical accuracy or efficiency. We also investigated case ordering, hypothesizing that forcing students to focus on structural characteristics without the support of non-diagnostic surface characteristics might ultimately immunize them from distraction by surface characteristics and improve their performance. Lastly we evaluated whether writing case summaries would improve performance on analogous cases.

Though results from this study were mixed, there was not strong evidence that applying explicit comparison of cases using ATE had a positive influence on measures of diagnostic performance. Likewise case ordering did not have a reliable effect on test case performance. However, the positive interaction seen between case ordering, ATE, and diagnostic category (i.e. cardiology versus bioterrorism) with respect to correct treatment score may suggest that explicit comparison can improve performance when students encounter difficult cases first, particularly when they are somewhat familiar with the basic principles involved in the diagnostic category. One explanation for this result is that explicit comparison of cases tends to focus students on all salient characteristics of the case – both diagnostic structural characteristics and non-diagnostic surface characteristics. In the "easy" case ordering, surface and structural characteristics were aligned across early cases and thus students may have mistakenly associated surface characteristics with diagnostic efficacy. In contrast, when non-diagnostic surface characteristics do not align with diagnostic structural characteristics in early cases, explicit comparison via the ATE seems to improve later treatment performance. This appears to be particularly true for cases in which the students may already be somewhat familiar with the diagnostic domain (i.e. cardiology rather than bioterrorism cases). Specifically, for cases in which the students are more familiar with the treatment principles, ATE instructions focus them on structural comparisons that the "hard" case ordering encourages. This suggests that for new areas of medical learning it is necessary to first educate students on important treatment principles before moving on to case-based learning methods.

One reason why this interaction was seen on correct treatment scores, and not correct diagnosis scores, may be that medical students tend to rely more heavily on causal reasoning when generating differential diagnoses. It is possible that medical diagnosis for novice medical students may not be principally analogical in nature, but rather driven by pathophysiological correlations (see [13]). However, once a diagnosis is correctly determined, relevant exemplar cases may be useful for determining treatment analogically. This explanation would then make it unlikely that ATE failed to improve subjects' case performance because medical students (in contrast to Gentner, Lowenstein and Thompson's business students [8]) instinctively think analogically and therefore forcing explicit comparison of cases is superfluous. It is more likely that the students participating in this study did not have the necessary expertise to abstract the diagnostic principles from the cases. Thus, explicit comparison of cases did not reinforce these principles (see Chi, Feltovich, & Glaser [14] for a similar issue in the domain of physics). In contrast, expert clinicians may rely on a repertoire of cases, or 'illness scripts', built from personal clinical experience when applying diagnostic reasoning [15].

There are other potential reasons why ATE did not have a more significant effect on overall case performance. It may be that subjects' memory of analogs simply decayed over time – a per-subject analysis of case performance against time between study initiation and study completion would be complicated and has not been attempted. While this explanation is possible, medical professionals obviously use very old knowledge from previous cases to diagnose new cases. Thus, this explanation may again interact with the experience level of the student/professional.

Lastly, explicit comparison may have failed to produce a greater effect because of the complexity and interactive nature of the case analogs themselves. Previous studies on the effect of analogical reasoning have used relatively simplistic case scenarios with fewer variables to consider and categorize as superficial versus structural characteristics [e.g. 13]. Also, in those studies, information was passively transferred, such that subjects were assured of being exposed to all structural characteristics germane to solving the problem. In contrast, cases which offer comprehensive detail, such as the medical cases used in this study, may present too many variables to make definite determinations about the similarities and differences between cases. Subjects could be overwhelmed when trying to determine which structural categories of data (historical, diagnostic, therapeutic, etc.) are most important to compare and contrast. Furthermore, due to the highly interactive nature of ***Medulator*** cases (emulating real-world information gathering), subjects may fail to elicit certain critical structural information required for source analog comparison.

There were some notable limitations to the present study. Human factors could have led to overestimation of some dependent variables in all groups. For example, when selecting diagnostic tests, correct diagnoses, or correct treatments, subjects could potentially take a "shotgun approach" by selecting more options than necessary in order to observe the feedback or to ensure a correct selection, leading to spuriously high numbers of incorrect selections.

Another potential limitation was the variability in the time period over which subjects completed cases. Subjects were initially urged to complete all cases within 2 weeks. Previous laboratory studies requiring analogical reminding have typically been conducted over one week or less [9]. Justification for this stipulation is founded in the concept that analogical reasoning requires the subject to be able to recall the pertinent structures of known cases. Case summaries served this purpose to the extent that subjects were insightful and diligent in their self-generated accounts of each case. Nevertheless, when excessive periods of time transpire between the source case and the target case, an unpredictable degree of information decay can occur, limiting the usefulness of the subject's recall. However, in order to reach the target of subject completions (and avoid participant drop-out), study account extensions were granted in one-week increments. Thus, between-subject variance may have dramatically increased because some students completed all cases within a brief period of time (e.g. 24 hours) versus others who took more than a month to complete all 11 cases, resulting in weeks between their first and last cases.

Of note, and of particular interest to designers of patient simulation programs, is case performance improvement resulting from the use of case summaries as a tool for reflection and as a proxy for self-explanation. In the present study, diagnostic accuracy improved and treatment attempts were reduced in participants who summarized in writing the pertinent information in each case. This effect was observed early (from the first case in each sequence) and was sustained throughout each case sequence. While this result is consistent with previous research [1-4], and in itself is not surprising – after all, case summaries are analogous to physicians writing an "H&P" or a "SOAP note" – these investigators are aware of no comparable computer-based patient simulation platforms which incorporate such a simple

yet effective tool. Furthermore, informal analysis of subjects' case summaries written for this study revealed that even subjects in groups which did not perform ATE (and, therefore, did not need case summaries as a recall tool) wrote word counts comparable to ATE subjects, suggesting that those subjects were using the case summaries for reflection and self-explanation.

Finally, the follow-on survey results demonstrate a high degree of user acceptance of ***Medulator***. Subject experience with patient simulation tools such as ***Medulator*** is largely lacking, with 65% of subjects responding that they have used similar tools only rarely (once per year or less) or never. Although the cases in this study operated in pure assessment mode, user comments indicated that many considered ***Medulator*** an excellent learning tool—an observation supported by their improvement in performance during the course of study. However, subjects perceived explicit analogical reasoning to be more effective than it was. Nearly all (97%) of ATE subjects felt that ***Medulator*** was at least somewhat effective at helping them to understand the concept of analogical (case-based) reasoning. Three-fourths of ATE subjects also felt that analogical reasoning was at least somewhat helpful to them in solving cases, even though objective measurements did not agree (ATE was not statistically associated with improved diagnostic accuracy). Interestingly, the applicability of bioterrorism cases to subjects' training was rated low.

In conclusion, we found that the modifications made to ***Medulator*** for manipulating case sequences and supporting explicit case comparison yielded mixed results. There may be evidence that such conditions have a greater effect on therapeutic reasoning than diagnostic reasoning, primarily when applied in familiar knowledge domains. However, using case summaries as a tool for reflection and proxy for self-explanation led to significant improvement in students' performance. It is likely that level of expertise is a primary determinant of whether clinicians use causal versus analogical reasoning in the overall diagnostic reasoning process.

6. References

1. Holyoak, K.J., & Thagard, P. (1995). Mental leaps: Analogy in Creative Thought. MIT Press: Cambridge, MA.
2. Gentner, D., Holyoak, K.J., & Kokinov, B.N. (eds) (2001). The analogical mind: Perspectives from cognitive science. MIT Press: Cambridge, MA.
3. Holyoak, K. J. (2005). Analogy. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 117-142). Cambridge, UK: Cambridge University Press.
4. Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, *67*, 2797-2822.
5. Goldstone, R.L., & Sakamoto, Y. (2003). The transfer of abstract principles governing complex adaptive systems. *Cognitive Psychology, 46,* 414-466.
6. Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1-38.
7. Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 1147-1156.
8. Gentner, D., Lowenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology. 95,* 393-408.
9. Bielaczyc, K., Pirolli, P. L., & Brown, A. L. (1995). Training in Self-Explanation and Self-Regulation Strategies - Investigating the Effects of Knowledge Acquisition Activities on Problem-Solving. Cognition and Instruction, 13(2), 221-252.
10. Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. Cognitive Science, 5, 145-182.
11. Neuman, Y., & Schwarz, B. (1998). Is self-explanation while solving problems helpful? The case of analogical problem-solving. British Journal of Educational Psychology, 68, 15-24.
12. Chi, M. T. H., Deleeuw, N., Chiu, M. H., & Lavancher, C. (1994). Eliciting Self-Explanations Improves Understanding. Cognitive Science, 18(3), 439-477.
13. Patel, V.L., Arocha, J.F., Zhang, J. (2005). Thinking and reasoning in medicine. In K.J. Holyoak & R.G. Morrison (Eds.) *The Cambridge Handbook of Thinking and Reasoning*. New York: Cambridge University Press.
14. R Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121-152.
15. Sisson, J., Donnelly, M., Hess, G., & Woolliscroft, J. (1991). The characteristics of early diagnostic hypotheses generated by physicians (experts) and students (novices) at one medical school. *Academic Medicine*, 66, 607-612.