Autonomous Virtual Patients in Dentistry: System Accuracy and Expert Versus Novice Comparison

Glenn T. Clark, M.S., D.D.S.; Apar Suri, M.S.; Reyes Enciso, Ph.D.

Abstract: This article examines an autonomous virtual patient (AVP) system for identifying differences between novices and experts in dentistry. The two groups in the study were ten boarded or board-eligible experts (seven males, three females; mean±sd age 40±11) and twenty-six fourth-year dental students (fifteen males, eleven females; mean±sd age 27±3), who were defined as novices. All participants interviewed and mock-examined four randomly selected AVPs who had either orofacial pain or an oral medicine problem; they then selected needed diagnostic tests, diagnoses, treatments, and medications. The mean misrecognition rate of the software was between 13 and 19 percent. Data collected were examined for a difference between the two groups (novices versus experts) on multiple variables. Significant group differences existed in the final total score, the number of diagnostic tests ordered, and the number of medications selected. Novices reported that they found virtual patients to be a valuable educational experience. These data demonstrated that experts and novices asked essentially the same questions and spent similar amounts of time with the patients, yet the experts consistently scored higher and ordered fewer diagnostic tests and medications than the novices.

Dr. Clark is Professor and Director of Orofacial Pain Clinic, Division of Periodontics, Diagnostic Sciences, and Dental Hygiene, Herman Ostrow School of Dentistry, University of Southern California; Apar Suri is Programmer Analyst II, Center for Creative Technologies, Institute for Creative Technologies, University of Southern California; and Dr. Enciso is Assistant Professor of Clinical Dentistry, Division of Dental Public Health and Pediatric Dentistry, Herman Ostrow School of Dentistry, University of Southern California. Direct correspondence and requests for reprints to Dr. Glenn T. Clark, Herman Ostrow School of Dentistry, University of Southern California, DEN B14, 925 West 34th St., Los Angeles, CA 90089; 213-821-5298 phone; 213-740-3573 fax; gtc@usc.edu.

Keywords: dental education, educational methodology, computer-assisted instruction, virtual patients

Submitted for publication 4/20/11; accepted 12/30/1

Virtual patients (VPs) have a very specific purpose: to provide training for novice health care practitioners without risk to the patient. VPs also allow for delivery of a consistent set of patient cases to the novice-in-training, something a clinical rotation cannot do. For the most common type of VP, a user progressively gathers information about a specific patient by clicking on a series of linked information pages before making a final decision about the patient’s diagnosis and treatment. This type of system has value in that it can introduce preclinical health care students to the process of history and examination data-gathering and give them familiarity with the decision-making process. Unfortunately, with this linked-page style of VP, student users do not formulate their own questions, which is an essential skill that they must acquire. At most, it forces the student user to choose pre-scripted questions and remember the resulting information before he or she makes choices. Another type of VP system is one in which student users (by speech or text) submit their own questions and the VP recognizes and responds with an appropriate answer. This type of VP is called an autonomous virtual patient (AVP) and is the focus of this article.

The underlying software for an AVP is a natural language recognition program that can identify and respond to user-generated queries. Such systems clearly have the potential to be immersive and highly engaging as they are refined. Prior research has found that standardized patients (SPs) receive higher ratings with respect to empathy, immersion in the scenario, and several nonverbal communication responses. However, interaction with the VP was still considered acceptable as a learning experience and had the added value of repetitive practice possibility. In fact, the student users rated their AVP experience as 7.2±1.8 (on a 1 to 10 scale with 10=highest), compared to a rating of 7.5±1.2 for their interaction with SPs. To some degree, the reality gap may have been related to the fact that, for this VP system, student user questions were correctly recognized and responded to only 60 percent of the time. This recognition rate was similar to that in another VP report that dealt with traumatic stress disorders.
One study found that prior experience with SPs made medical students less likely to be rejected as acceptable by their actual patients. The SPs also made the students feel more comfortable during the examination and five times less likely to fail the examination. To date, no data exist on the long-term effects of AVP-style medical encounter training as a student on actual communication ability later in practice.

Overall, the literature suggests that while AVPs as they currently exist are not as engaging as a live standardized human patient, the interaction is still acceptable as an educational experience. The biggest flaw with an AVP system is that the student user questions are frequently not recognized and so must be restated or redirected. Of course, as natural language recognition software improves, such issues will become less important. Also, training with an SP in medical school does not necessarily predict high communication skills in a residency program. The purpose of our study was to introduce and describe an AVP educational system for dentistry and to compare experts and novices during an AVP encounter. The primary null hypotheses were that there are no differences between the experts and novices in 1) number of questions asked; 2) number of diagnostic tests ordered and medications prescribed; and 3) final total score received on these cases.

Methods

The two subject groups in this study were a group of orofacial pain and oral medicine specialists (defined as experts) and a group of fourth-year dental students (defined as novices) attending a clinical rotation in the orofacial pain and oral medicine center, where they could observe care provided by residents. At the end of the rotation, the students were asked to interview a set of AVPs presenting with one of several common conditions that might be seen in private practice. The experts (seven males and three females; mean age 40±11; range 29–61 years) were all boarded or board-eligible specialists in the fields of orofacial pain and oral medicine. Two of the eight experts were boarded only in orofacial pain and oral medicine, and one was boarded in oral pathology only. The novices (fourteen females and twelve males; mean age 27±3; range 24–35 years) were all students in good standing in the fourth year of a four-year D.D.S. program.

Eight individual cases from the patient pool were created for this study. The cases involved the following diseases: clenching induced myalgia, arthralgia, and otalgia; generalized osteoarthritis affecting the TMJ; acute traumatic trismus of the jaw; acute disk displacement without reduction; chronic radiation induced sialoadenitis; lip mucocle; erosive lichen planus; and squamous cell carcinoma of the mouth. The novices and experts randomly chose four cases as follows. Each user was presented with eight slips of paper (representing cases one through eight) that had been shuffled and turned over. The user selected four slips, and those were the cases assigned to that novice or expert.

All subjects were instructed about how to use the AVP system. They were told that they were to ask medical interview questions as needed, then to ask physical examination questions as needed, and finally to select the final choices for the case and try to do all of this within thirty minutes. All participants were provided with a one-page guidesheet that outlined the common elements of a medical interview and physical examination of the head and neck region.

AVPs require an underlying natural language software program, which was custom written in C# programming language. This program has four elements: input/send, tagging, comparison/best match, and response. The input process involves capturing text or speech input from users and sending it to a server that hosts a natural language-processing program. In the tagging element, all input data undergo disambiguation and tagging of the words in the data set. Once the data are tagged, they are then compared to a set of tagged pre-existing input, specifically the questions in the database of probable questions. This comparison involves checking for syntactic and semantic similarities between the user input and pre-existing questions to determine the best match. In the response element, the best-matched question has a linked answer in the database that is then sent as the reply. This process is called natural language recognition and is central to creating an AVP.

A database of probable questions and answers must be established before any natural language program will work. To create the AVPs, it was necessary to craft 178 unique question-answer categories to the common medical interview questions and 138 question-answer categories to the common head and neck examination items performed by a specialist. Since there are many ways of saying the same thing, there are always multiple questions within each question-answer category although there is...
only one answer. For example, “What is your chief complaint?” or “Why did you come in today?” or “Why are you here?” are all questions that would result in a single answer—namely, the patient’s chief complaint. For the questions dataset, one investigator (GTC) created a database of 1,053 interview questions and 738 physical examination questions (approximately five questions for every answer). The animation routines are not sophisticated enough to have the user’s avatar actually examine the face and mouth of the AVP; instead, the examination process used in the system required the user to state exactly (using correct anatomic terminology) what aspect of the examination he or she wanted to perform. Once asked, the appropriate answer to this examination would be returned by the AVP. For example, saying “examine the deep masseter muscle for tenderness” would produce an answer that described the results of this palpation. In this set of patients, images of the oral tissues were not provided. Instead, we relied on verbal descriptions of all abnormalities. With regard to creating a new AVP, it is not necessary to change all the answers since many of the answers will be negative and will not change from patient to patient. Approximately one-third to one-half of the answers will need to be rewritten when creating a new AVP. The question set does not change very much, if at all, from patient to patient since it is mostly independent of the answers.

The AVP interface could be alternating lines of text with the doctor’s question followed by the patient’s answer. However, to make the interface more interesting and more engaging, this research group elected to host the AVP in a virtual world environment. Specifically, Second Life (www.secondlife.com) was selected, and a clinic was created where the eight VPs were located (Figure 1). The users logged in with preassigned three-dimensional avatars. Under the user’s keyboard control, these avatars were able to walk from room to room, could interview and examine each patient, and then log on to a final choice screen where they made the final choices.

The final choices were divided into four sub-screens; the oral medicine cases and the orofacial pain cases had different choices on these screens. Both had four sequentially available sub-screens:

Figure 1. The Second Life island (A) where the eight-chair clinic was constructed (B; top view); the autonomous virtual patient seated in the dental chair (C); and a dentist’s avatar seated in front of the patient interviewing her (D)
Results

For the current AVP system, the mean (±1 s.e.m.) misrecognition rate was 14±1 percent for the medical interview questions portion and 13±2 percent for the physical examination portion of the encounter by the experts. The rate was 16±1 percent for the medical interview questions portion and 19±2 percent for the physical examination portion of the encounter by the novices.

The interview and examination questions asked, the final choices selected, the total standardized final score, and the elapsed encounter time for ten experts and twenty-six novices for the cases evaluated are shown in Table 1. There were no significant group differences between novices and experts with the exception of the total normalized final choice score, the number of diagnostic tests ordered, and the number of medications selected.

The main analyses compared the two groups (novices versus experts) with regard to the number of questions asked (both interview and examination); the number of diagnostic tests and medications ordered; and the final total score received. Because some cases were harder than others and the maximum number of points possible varied by case, the scores were normalized to a percentage of the maximum score achieved on the case. In regards to accuracy, the transcript of the actual questions asked was compared to the transcript of the question-answer set with which it was matched. No variables were normally distributed, so the group comparisons were performed using the nonparametric test for independent measurements (Mann-Whitney U test). All calculations were performed using SPSS 12.0 for Windows software; p-values of less than 0.05 were reported as statistically significant.

This research met the requirements outlined in 45 CFR 46.101(b)(4) and qualified for exemption from Institutional Review Board review. This exemption was approved by the University Park Institutional Review Board of the University of Southern California (USC UPIRB # UP-11-00292).

Discussion

This study compared the encounter length, number and type of questions asked, and final choice preferences for a group of ten experts in orofacial pain and oral medicine with a group of twenty-six novice dental students. Both groups worked with a set of AVPs with various diseases that would appear in a specialty clinic dedicated to orofacial pain and oral medicine. No differences between novices and experts were found with regard to the number of interview and examination questions asked and the encounter time. In addition, the two groups were not significantly different with respect to the total number of final choices selected.

Specifically, the experts and novices utilized approximately twenty-three minutes on average to complete their work-up of their AVP. The experts asked an average of one question every forty-five to forty-eight seconds. This speed is slightly slower than what occurs in an actual patient encounter (one per thirty and
one per forty seconds), according to the published literature.\textsuperscript{10-14} This slower rate of asking questions in the AVP scenario might be partially explained by the time it takes to type the question and the six-second delay for the software to recognize and answer the question. While this delay slightly interferes with the flow of dialogue between the patient and doctor, it was not, in our opinion, an important interference.

Significant group differences were found for the final choice score, the number of diagnostic tests ordered, and the number of medications prescribed. As would be expected, the experts scored significantly higher than the novices on the final score, and they ordered fewer diagnostic tests and medications than the novices. One can speculate that novices are more likely to order diagnostic tests, hoping they will provide valuable information about the case’s diagnosis, and they have little familiarity with medications so will misuse them. Of course, inappropriate diagnostic tests and medications are both expensive and increase the risk of false positive results and adverse drug reactions. By multiplying the rate of incorrect diagnostic choices for the two groups by the number of choices made, the data in this study show that the novices were 2.2 times more likely to order an inappropriate diagnostic test than the experts and 2.3 times more likely to prescribe inappropriate medications. Due to the relatively small sample size, the study was underpowered to find a significant difference in number of treatments selected per case.

Misrecognition of a question from a user is one of the biggest reasons the user feels that an AVP
system is not fully believable. Of course, questions can be misrecognized because a user badly mistypes the question, the user uses text speech (e.g., “what R U doing here?”), or the user asks a question out of the expected domain of a medical interview and clinical examination process (e.g., “have you seen a good movie lately?”), or the system is not fully trained. As the system is trained, questions are added to the database of probable questions, and the system is enhanced with classifier software and recognition algorithms, the accuracy will improve to an acceptable level (<5 percent misrecognition). For the current system, the accuracy rate was between 12 and 19 percent. The misrecognition rate in this study was lower than has been reported for other AVP systems, but this might be because each user was given a two-page guide with a set of sample medical interview questions and examination topics to consider using. Mismatched answers were not greatly problematic since most of the users were able to continue with questioning and receive appropriate responses to perform a diagnosis and perform treatment planning.

The novice questionnaire data from our pilot study suggest that the response delay and the question-answer mismatch were not a great deterrent in the user’s acceptance of the system. In fact, 40 percent of the student users strongly agreed that the VP dialogue was reasonably realistic, with another 41 percent slightly agreeing to this view. It is expected that the AVP system described in this article will improve as the speed of the system increases and the question-answer mismatch rate decreases. Two clear gaps in the system are that the AVPs do not ask questions of their doctor and the physical examination is not animated. While AVP systems with more realistic examination animations and AVPs that can ask questions are possible, those elements were beyond the scope of the system described in this study.

On the positive side, AVPs have the advantage over SP actors that the patients are absolutely consistent from user to user, and if enough cases are created, students may practice and hone their interviewing skills around the clock with these AVPs in preparation for their VP or SP objective clinical examination. Of course, SPs are sometimes used for training purposes, but their availability is limited and they are mostly used for testing purposes. The remaining question is whether the student user would improve in the real world with regard to his or her communication skills and his or her ability to establish rapport with the patient as a direct result of using an AVP system.

REFERENCES