Training Otologic Surgical Skills Through Simulation—Moving Toward Validation: A Pilot Study and Lessons Learned

Gregory J. Wiet, MD
Jeff C. Rastatter, MD
Sumit Bapna, MD
Mark Packer, MD
Don Stredney, MA
D. Bradley Welling, MD, PhD

Abstract

Introduction Methods for surgical education and training have changed little over the years. Recent calls to improve surgical efficiency and safety impose additional pressures that have an impact on surgical education and training.

Use of Simulation Integration of data from advanced imaging technologies and computer technologies are creating simulation environments of unprecedented realism. Surgical education and training are poised to exploit low-cost simulation technologies to mitigate these pressures that are having an adverse impact on curricula. To become effective, simulation needs to undergo rigorous validation studies.

Intervention With funding from that National Institute on Deafness and Other Communicative Disorders, we have embarked on a research design project to develop, disseminate, and validate a surgical system for use in otologic resident training and assessment and present key steps from this process.

Discussion We discuss limiting factors related to technology and conducting multi-institutional studies, along with current developments to integrate curricula, as well as training and assessment capabilities in surgical education using simulation.

Introduction

Over the years, surgical technique has evolved into a wide range of approaches to provide safer, less invasive, and effective treatment for a range of human ailments. As sophistication increased and outcomes improved, the perception of surgical intervention became viewed by the general public as safe and efficient and, at times, preferred to medical management because of its more rapid and direct approach to disease resolution. Although surgical techniques have undergone refinement, the methodology for surgical education has not. The methodology of anatomic study and observation of procedural technique was established by Halsted and Osler[1] at Johns Hopkins in the late 1800s. Hands-on "practice" began with simple procedural technique and progressively increased in complexity, first on cadavers and culminating with live patients. This incremental approach is the "gold standard" used today.

In addition, methods to evaluate a surgeon's technical skill are largely unchanged, remaining subjective and nonuniform. In most cases (including national specialty board examinations), readiness for surgical practice is determined by testing the fund of knowledge through written and open-ended oral examinations. Candidates for certification submit a list of the "appropriate" number of procedures performed in training. The assumption is that with the proper fund of knowledge and an "appropriate" number of procedures, individuals are assured to be technically skilled. Certification of technical expertise is based on the subjective opinion of program directors, who ultimately approve requests for surgical privileges.

Recent developments have had a considerable impact on surgical training and evaluation. The Institute of Medicine's report on medical errors called for safeguards to improve health care.[2] The report inevitably led to new pressures on

Gregory J. Wiet, MD, is from the Department of Otolaryngology–Head and Neck Surgery, The Ohio State University Medical Center; the Department of Otolaryngology, Nationwide Children's Hospital; and the Ohio Supercomputer Center and the Department of Biomedical Informatics, The Ohio State University; **Jeff C. Rastatter, MD,** is from the Department of Otolaryngology–Head and Neck Surgery, The Ohio State University Medical Center; and the Department of Otolaryngology, Nationwide Children's Hospital; **Sumit Bapna, MD,** is from the Department of Otolaryngology–Head and Neck Surgery, The Ohio State University Medical Center; and the Department of Otolaryngology, Nationwide Children's Hospital; **Mark Packer, MD,** is from the Department of Otolaryngology–Head and Neck Surgery, The Ohio State University Medical Center; and the Department of Otolaryngology, Nationwide Children's Hospital; **Don Stredney, MA,** is from the Ohio Supercomputer Center and the Department of Biomedical Informatics, The Ohio State University; and **D. Bradley Welling, MD, PhD,** is from the Department of Otolaryngology–Head and Neck Surgery, The Ohio State University Medical Center.

Portions of this work were presented at the poster session of the 2007 Annual American Academy of Otolaryngology–Head and Neck Surgery Foundation Meeting, Washington, DC.

This work was supported by National Institutes of Health/National Institute of Deafness and Other Communication Disorders grant RO1 DC006458-01.

Corresponding author: Gregory J. Wiet, MD, FACS, FAAP, Department of Otolaryngology, Nationwide Children's Hospital, 700 Children's Drive, Columbus, OH 43205, 614.722.3856, Gregory.wiet@nationwidechildrens.org

DOI: 10.4300/01.01.0010

surgical training. With the advent of standards to limit duty hours, residents find less time for clinical practice. With increasing pressures to maintain efficiency and reduce risk, attending surgeons limit their ''teaching'' moments. The effort to establish standards has direct and profound ramifications on resident education. Simulation has the potential to mitigate several limitations presented by the current learning environment, including limitations of time, availability of resources, and expert tutoring, as well as the capability to provide more precise tracking of proficiencies and safety practices prior to clinical performance.

Simulation

Computed tomography and magnetic resonance imaging have led to the capture of unprecedented structural representations of human anatomy and to new standards in both diagnosis and the design of treatment plans. Integrated with digital computer hardware and advanced computer graphics and interface software, these data sets have provided the basis for progressively sophisticated simulation environments. This sophistication includes the emulation of complex anatomic representations, as well as increasingly intricate representations of procedural technique and interaction. The recent emergence of graphical processing units driven by the gaming industry provides the low-cost hardware and open-source software basis to achieve unprecedented simulations in a cost-effective system.

Similarly to adoption by the military and the aviation industry, medicine is ready to exploit the capabilities of simulation in training and assessment. Simulation includes deliberate practice in a nonthreatening environment, increased availability by providing on-demand sessions, and increased variance from potentially unlimited specimens or subjects. These systems provide increasingly realistic simulations, with more continuous and quantitative assessment, as well as the ability to provide true ''objective'' assessment using standardized models.

A prototype was developed under the R21 mechanism (National Institutes of Health/National Institute of Deafness and Other Communication Disorders [NIDCD] 1-R21 DC04515-01) and was found to have sufficient potential for extensibility.[3] The current R01 effort (NIDCD R01 DC006458-01) includes dissemination of open-source software to promote the acquisition and integration of the simulation environment. We describe the iterative development, improvement, and dissemination of simulation to characterize its efficacy in the otologic surgery curriculum.

System Description

The system used in the ongoing multi-institutional study was designed as an emulation of the cadaveric temporal bone laboratory. All aspects of the simulator were set and locked in January 2008 at the start of the multi-institution study. The system allows multiple representations of temporal bone specimens acquired from cadaveric specimens by employing a modified clinical protocol running on a 64-detector computed tomography scanner. Through direct volume-rendering techniques, the system provides arbitrary orientation of each specimen, allowing the user to determine the surgical approach. Tools are emulated through the use of a PhanTom OMNI haptic device from SensAble Technologies Inc., Boston, Massachusetts. The OMNI provides 2 operational buttons that can be accessed easily by the user's index finger. One allows for interactive orientation of the data around a central point; the other toggles the drill on or off. Closely associated with the amount of force presented on the drill, we can modulate sampled sounds recorded during an actual temporal bone dissection. Pitch modulations allow correlation of the amount of pressure being placed on the drill, as well as provide an auditory signal as to the amount of resistance of the bone being drilled.

Two Studies Influenced the Design of the Multi-Institutional Study

Two local studies informed the multi-institutional study. The first study validated the Welling Scale 1 (WS1), a 35-item binary grading system developed at The Ohio State University (OSU) for grading dissected temporal bones as a measure of technical performance.[4] The second was a formative pilot study conducted at OSU to characterize the multicenter validation study and to test the robustness of the system.

Validation of the Welling Scale was needed before it could be used as a framework for determining the efficacy of the surgical system in the curriculum. In this study, 12 residents in otolaryngology performed basic mastoidectomies with a facial recess approach on 26 cadaveric temporal bones. Six independent raters (2 neurotologists, 1 neurotology fellow, 1 pediatric otolaryngologist, and 1 general otolaryngologist) scored the resulting dissections on 2 separate occasions using the WS1. Raters were blinded to residents' year of training. The κ statistic was calculated for interrater and intrarater reliability. Intrarater agreement was high. Although most interrater agreement scores were moderate, there was very high interrater agreement between the 2 expert neurotologists.

An additional analysis of these data was performed to evaluate the components that contributed to measurement error, using the WS1 for rating performance. Rater disagreement introduced only a small error into scores. It was concluded that the WS1 has a small measurement error with 2 raters (neurotologists) and 2 bones for each study participant.[5]

The formative evaluation pilot was a prospective, randomized, blinded trial in 12 study participants (6 otolaryngology residents and 6 medical students) with no previous temporal bone training who were assigned to 1 of

TABLE 1 **Comparison of Posttest Dissection Scores**

	Welling Scale Values, %[a]		
	Mean[b]	Range	SD
Traditional Group[c]	23	8.6–46	8.6
Simulator Group	17	0–40	8.5

[a] Welling Scale values are reported as a percentage of the total possible points (35) that could be attained with each dissection. For example, a raw score of 35 would be a Welling Scale value of 100%.
[b] Difference between groups significant by mixed-effect model, $P = .05$.
[c] For traditional and simulator groups together, n = 12.

2 training groups: simulator or cadaveric lab. Participants received a limited standardized pretest education comprising a 30-minute didactic lecture, including a handout derived from the Nelson temporal bone dissector.[6] Participants were then randomized to 2 test groups. The simulator group was given unlimited access to the virtual temporal bone dissection system for individual practice during 2 weeks. The traditional group was given access to 2 cadaveric temporal bones each for individual practice during 2 weeks in the temporal bone laboratory. To reduce the required number of cadaveric bones needed for the study, we limited the number of practice bones to 2, with unlimited access. All participants then dissected 2 cadaveric temporal bones for a posttraining measure. Two senior neurotologists evaluated the participants' performances using the WS1. Performance was compared between the 2 groups. The protocol was analyzed for feasibility and ease of execution. The computer system was evaluated for robustness, and modifications were made for its use in the multicenter study.

A mixed-effect model was used to analyze these data. Mixed-effect models incorporate analysis of both fixed factors (for example, traditional versus simulator group) and random factors (variability among measures within participants). Even though all participants were considered novices, level of resident education (medical student, PGY1, PGY2) and their initial simulator skills test were included in the model as fixed effects. Total practice time was initially included in the model; however, it was shown to be highly insignificant because of the limited number of participants and high variability for the time periods used by the participants in this study, and it was excluded from the final model. Average practice times for the 2 groups were not significantly different. The traditional group times were: mean, 2.5 hours; range, 2.0 to 3.5 hours; and SD, 0.56 hours. The simulator group times were: mean, 3.3 hours; range, 0.75 to 5 hours; and SD, 1.9 hours. Pairwise comparisons were done using Tukey adjustment.

Comment on the Overall Results

The pilot study was designed to delineate the issues surrounding the overall study design and identify factors that would have an impact on the planned multicenter trial. The results reported are not necessarily expected to provide data for definitive conclusions regarding the efficacy or validity of the simulator in temporal bone dissection training. Results should be considered in the context of the limitations listed above.

Overall, posttraining dissection scores were very low (mean percentages of 23 and 17 for traditional and simulator groups, respectively). The traditional group performed better than the simulator group. The difference was statistically significant based on the mixed-effects model, with $P = .05$. (TABLE 1).

Effect of Study Participant Education Level

The mean WS1 scores were highest for the PGY2 residents and lowest for the medical students. The difference between scores of the PGY1 and PGY2 residents was not statistically significant. However, the differences between residents (PGY1 or PGY2) and medical students were significant ($P = .05$ and $P = .02$, respectively). This would suggest construct validity.

Design of the Multi-Institutional Study

In the transition to the multi-institutional study, the first limiting factor discovered was the restriction on participants' time because of study participants' other demands. The high variability in practice times, especially for the simulator group, was thought to be due to issues with the computer system robustness (freeze-ups) and lack of motivation to complete the study on the part of volunteer participants. Limiting the study to 2 weeks was required to accommodate time constraints on the participants. A $100 award was offered for the highest score on the WS1 as a motivator to practice.

A second limiting factor was access to, and cost of, cadaveric temporal bone specimens. In validation studies, it is imperative that new methodologies be compared to traditional ones. Practice on cadaveric specimens is the "gold standard" for perfecting technique in temporal bone surgery, and the WS1 is based on final product analysis of dissected cadaveric temporal bones. Ideally, study participants would drill bones to establish their

performance prior to randomization into the 2 training arms. These "pretraining" dissections could then be compared to "posttraining" dissections. Therefore, the training methodology that had greater improvement in performance would be the better methodology. This would require 6 cadaveric temporal bones (2 pretraining, 2 practice, and 2 posttraining) for participants randomized to the traditional arm, and 4 bones (2 pretraining and 2 posttraining) for those in the virtual training arm. To reach the appropriate statistical power, it was calculated that 50 participants in each arm (and a total of 500 cadaveric bones) would be needed to demonstrate a difference.

Many programs expressed interest in participating but did not have access to or could not afford the cadaveric temporal bones. This proved to be the single most challenging component to center recruitment, and it influenced study design. We eliminated the pretraining bones to run the pilot study (mentioned above). A major criticism of this approach is the lack of pretraining dissections as a measure of baseline performance that can be compared with posttraining dissections. This would be a valid and significant criticism. However, Fernandez et al[5] demonstrated that residents who had not previously had a formal temporal bone course (PGY1 and PGY2) had such low WS1 scores compared with those who had completed a formal temporal bone course that their starting performance measure was so close to 0 as to be statistically insignificant. Based on this analysis, the pretraining score for novice participants was assumed to be 0. For the expanded multicenter trial to recruit all training years, a pretraining virtual bone dissection was added along with 2 posttraining virtual dissections to provide both pretraining and posttraining measures of performance. By significantly reducing the number of required bones, we are able to recruit more study participants from multiple centers to reach our projected recruitment of 50 participants in each arm.

Time for deliberate practice and availability of resources as limiting factors illustrate the negative influences on the current training methods and barriers to research that seeks to evaluate different methodologies. Simulation technology may serve to mitigate these deficiencies by obviating the need for physical material in early training and supporting deliberate practice on demand.

The institutional review board at OSU reviewed and approved the human participant protocol. Participants were assured their performance data were protected from exposure, and they were instructed to contact the lead author if they perceived any coercion to participate in the study. The NIDCD required that establishment of a Data Safety and Monitoring Plan to oversee the project was registered on the ClinicalTrials.gov website because it was considered a phase 3 clinical trial by NIDCD officials. Participants were recruited and informed consent obtained according to approved protocol guidelines.

Overcoming Computer System Problems

The initial virtual system for the local trial experienced technical problems, which affected the outcomes of the pilot study. Initially, techniques for randomizing participants, assigning unique identifiers to each participant, and tracking which practice and testing bones were assigned to which participant were inadequate. Subsequently, software was written that allowed the local investigator to enter in specific information. The software would then handle randomization and assignments and output these into a readable or printable file. It became apparent that the computer environment was quite complex for individuals without knowledge of the software's design. Subsequent improvements included use of a button on the dexterous device (OMNI) to control orientation of the virtual specimen. These and other issues, such as the initial system's inability to accommodate stereo viewing, were taken into account in modifying the system for use in the multicenter trial.

Current Multi-Institutional Studies

The current multi-institutional study employs a modified human participant protocol and updated system, with residents in otolaryngology at all levels of training as participants. Each institution has the OSU institutional review board approved locally (see TABLE 2 for institutions involved). The goal of the study is to accrue 100 study participants (50 in each training arm).

As part of the overall research plan, the system has continued to be upgraded. Two dexterous interfaces, one for suction/irrigation and the second for drilling, have been added. Instrumentation is more elegantly displayed, with shadows that contribute cues that facilitate localization of tools. The new system also allows for the introduction of bleeding effects. This imparts a sense of consequence and can be linked with errors, such as improper attention to cleaning the wound and/or hitting critical vasculature. More recent enhancements will be used in future trials.

Discussion

Additional studies continue to demonstrate the effective use of simulator training in a number of areas. Use in training for laparoscopic surgery has been the most established application of this technology. A recent study found that, "In people with no laparoscopic experience, virtual reality (VR) training is better than no training in relation to the time taken to perform a task, improving accuracy and decreasing error…"[7] Another recent review outlined the current use of systems in laparoscopic, urologic, bronchoscopic, sinus, and other areas of development.[8] The key to effective use of simulators is not only in the development of simulators but, more importantly, in how they will be used (curriculum development) and how results of training will be measured (skills assessment). A recent review noted that there is a considerable amount to accomplish for simulation to be

TABLE 2 **Institutions Enrolled in the Study**

Institution	Status	No. of Participants
Duke University	Completed	6
Massachusetts Eye and Ear Infirmary, Harvard Medical School	In process	...
University of Texas, Southwestern	Completed	10
University of Iowa	Completed	15
Baylor College of Medicine	In process	...
University of Mississippi	Completed	8
Henry Ford Hospital System	Completed	8
Eastern Virginia College of Medicine	In process	5
Albert Einstein College of Medicine, Montefiore Medical Center	In process	...
University of Cincinnati	In process	...
Virginia Commonwealth University	In process	...

... means subjects not yet recruited.

accepted as an integral part of surgical training, specifically in the area of curriculum development and the acquisition of cognitive knowledge along with hands-on skills.[9] Future goals include development of a standardized curriculum that integrates the temporal bone simulator with an enhanced cognitive learning process that responds to the need to identify and manage errors through active feedback to the trainees. Confirmation that the simulation environment shortens the learning curve is essential. Development of a common language for measurement of technical performance is essential to demonstrate the efficacy of simulation in surgical training.

Our report outlined the OSU experience with the development of a "virtual cadaveric temporal bone laboratory" for mastoidectomy surgery based on simulation technology. Rapidly advancing technologies in surgical simulation continue to allow for the evolution of temporal bone dissection simulators with more realistic features. During the past 5 years, our group has followed an iterative design, implementing numerous modifications in our temporal bone dissection simulator that aim to make the simulated environment more similar to the natural environment of both cadaveric dissection laboratory and actual surgery. Advances in graphics acceleration hardware and computer performance allow smoother interactions in real time. Sensitive haptic devices provide tactile feedback, and high-resolution displays provide realistic visual interfaces during simulated surgery. High-resolution micro computer tomography (μCT) scanners and high-field (7-T) magnetic resonance imagers are providing volumetric image data sets with superior anatomic detail.[10]

Conducting a local pilot study proved to be valuable before extension to a multi-institutional study. Limitations of the current training paradigms (time and material) had a significant impact on study design and proved to be limiting factors in ensuring the scientific rigor necessary to validate the simulation. It is imperative that this type of study continue to be performed and further refined to provide a scientifically rigorous evaluation of the usefulness of simulation in technical skills training and assessment. As mentioned above, integration of these systems into current curriculum is necessary to leverage their true potential. We plan a future project with integration of a curriculum for teaching otologic surgery in a consortium of active centers. Additionally, we plan to demonstrate the advantage of simulation for technical skills assessment within our consortium by developing widely accepted metrics applicable to a simulation environment.[11] The application of simulation technology to surgical skills training is consistent with recent Accreditation Council for Graduate Medical Education Committee recommendations for innovation and improvement in the learning environment.[12] As these methods are developed and refined, the importance of solid scientific rigor associated with testing of efficacy cannot be overstated. Because this is a new area for surgical training, time is needed to develop systems, curricula, and integration into training programs. This time will, however, be well spent, and individual acceptance of this type of training is necessary to realize the true advantages of simulation training and assessment.

References

1 Osborne MP. William Stewart Halsted: his life and contributions to surgery. *Lancet Oncol*. 2007;8(3):256–265.

2 Kohn LT, Corrigan JM, Donaldson MS; Committee on Quality of Health Care in America, eds. *To Err Is Human: Building a Safer Health System*. Washington, DC: National Academy Press; 2000.

3 Bryan J, Stredney D, Wiet GJ, Sessanna D. Virtual temporal bone dissection: a case study. *Proc IEEE Vis*. 2001;497–500.

4 Butler NN, Wiet GJ. Reliability of the Welling Scale (WS1) for rating temporal bone dissection performance. *Laryngoscope*. 2007;117:1803–1808.

5 Fernandez SA, Wiet GJ, Butler NN, Welling DB, Jajoura D. Reliability of surgical skills scores in otolaryngology residents: analysis using generalizability theory. *Eval Health Prof*. 2008;31(4):419–436.

6 Nelson RA. *Temporal Bone Surgical Dissection Manual*. 2nd ed. Los Angeles, CA: House Ear Institute; 1991.

7 Gurusamy K, Aggarwal R, Palanivelu L, Davidson BR. Systematic review of randomized controlled trials on the effectiveness of virtual reality training for laparoscopic surgery. *Br J Surg*. 2008;95(9):1088–1097.

8 Satava R. Historical review of surgical simulation–a personal perspective. *World J Surg*. 2008;32:141–148.

9 Aggarwal R, Darzi A, Grantcharov TP. Re: A systematic review of skills transfer after surgical simulation training. *Ann Surg*. 2008;248(4):690–691; author reply 691.

10 Wiet GJ, Schmalbrock P, Powell K, Stredney D. Use of ultra high resolution data for temporal bone dissection simulation. *Otolaryngol Head Neck Surg*. 2005;133:911–915.

11 Wan D, Wiet GJ, Kerwin T, Stredney D, Welling DB, Dodson E. Objective assessment of temporal bone dissection. *ARO Abstr*. 2009;32:60.

12 Committee on Innovation in the Learning Environment. *Fostering Innovation and Improvement in the Learning Environment Through Accreditation*. 2007.