

The Journal of Education in Perioperative Medicine

BRIEF REPORT

A Randomized Controlled Pilot Study Comparing ChatGPT to Live Coaching Deliberate Practice to Teach Difficult Conversations

MERON AMARIW
MITCHELL L. PHILLIPS

CLARA S. MILLER
JEFFREY H. BARSUK

JOHN D. MITCHELL
HEATHER A. BALLARD

INTRODUCTION

Difficult conversations refer to clinical encounters in which clinicians must communicate information that may significantly alter a patient's view of the future, negatively impact quality of life, or evoke strong emotional responses.¹ These discussions require a balance of professionalism, clarity, and empathy. Poor performance in difficult conversations has been associated with both decreased patient satisfaction and increased clinician burnout.^{2,3} Communication skills are teachable and can be developed with simulated patients (SPs) and coaching.⁴ Artificial intelligence (AI) has also been heralded as a potential tool to teach a variety of medical skills, including communication.⁵ Whereas AI shows promise in improving patient–physician interactions, its effectiveness in preparing anesthesiology trainees for difficult conversations remains unclear. AI has the potential to improve communication skills at lower training costs by reducing the need for trained SPs and faculty. The primary aim of this study was to perform a pilot study comparing 2 difficult conversation training approaches, didactic instruction followed by deliberate practice using either OpenAI's ChatGPT (San Francisco, CA) or using an SP with live coaching. We hypothesized that a training approach combining didactic instruction with deliberate practice using ChatGPT would be comparable to deliberate practice

with an SP and live coach in improving communication skills.

MATERIALS AND METHODS

We performed a randomized controlled trial of ChatGPT-based deliberate practice communication sessions for rotating anesthesia trainees at Ann & Robert H. Lurie Children's Hospital of Chicago from April to June 2024 (Figure 1). The Lurie Children's Institution Review Board approved the study protocol before subject enrollment. Participants were recruited from rotating student registered nurse anesthetists and anesthesiology residents during the study period. All eligible rotating trainees were invited to participate, and no eligible trainees declined enrollment. After providing informed consent, participants were randomized into 1 of 2 groups: deliberate practice with ChatGPT (ChatGPT group) or deliberate practice with an SP and live coaching (SP group) by using a random number generator (<https://www.random.org/>).

Following randomization, each participant answered questions on their years of training, previous communication training, and number of times having difficult conversations. Subsequently, they underwent a pretest and were scored using a skills checklist.⁴ During this assessment, the participant was asked to disclose a medication error to an SP in which the participant had given the SP's child a known allergen for antibiotic prophylaxis.

Next, all participants attended a faculty-led didactic session focused on the SPIKES (setting, perception, invitation, knowledge, empathy, summarize)⁶ framework for difficult conversations. This didactic session was delivered to participants prior to deliberate practice with either ChatGPT or an SP.

Based upon their study assignment, participants underwent either deliberate practice with an SP and were given feedback from a VitalTalk instructor (HB) or deliberate practice with ChatGPT and received feedback from ChatGPT only. We conducted the study using the ChatGPT-3.5 language model (<https://chat.openai.com/chat>). Participants accessed the platform via generic, nonidentifiable accounts, and all scenarios were fictional with no use of protected health information (PHI), obviating the need for a Health Insurance Portability and Accountability Act–compliant environment or privacy gateway under our institutional review board determination. The initial input prompt in Supplemental 1 is an example of one of the deliberate practice scenarios (rocuronium overdose) designed to replicate the scenarios used in the SP group. These practice scenarios were intentionally varied and included the following clinical situations: rocuronium overdose requiring postoperative intubation and ventilation, ketolorac administration in

continued on next page

continued from previous page

a patient with underlying renal disease, and a failed caudal nerve block with poor postoperative pain control, allowing learners to engage in different variations of difficult conversations.

Participants in the ChatGPT group engaged in a written interaction with ChatGPT acting as the simulated patient and subsequently received coaching from ChatGPT based on the SPIKES framework. The written version of ChatGPT was utilized because voice-to-text capability was not available during the study period. Participants in the ChatGPT group were able to choose how long they practiced and their number of attempts with each scenario. Following deliberate practice, both the ChatGPT and SP groups underwent a posttest with a live SP and were scored using the same skills checklist as the pretest. The same scenarios were used to facilitate consistent scoring criteria and preserve evaluator blinding to assessment timing. After course completion, participants were surveyed on their course satisfaction, self-reported knowledge, confidence, and application before and after training on a Likert scale of 1 to 5 (1: *strongly disagree*; 5: *strongly agree*).

All testing sessions were video-recorded and stored using deidentified file names. Two trained and calibrated graders (MA and CSM), blind to pretest/posttest status, independently evaluated each video-recorded test. Graders were able to view both the standardized patient and the learner in the recordings. Graders used a previously developed 16-item skills checklist on which participants' performance scores were graded dichotomously as *done correctly* or *not done correctly* (Supplemental 2).⁴ The checklist structure was modeled on the SPIKES framework and mirrors previously published SPIKES-based instruments that have demonstrated high interrater reliability and generalizability (medical students, trainees, and faculty).^{4,7} Checklist items assessed key domains of difficult conversation performance, including rapport building, assessment of family understanding, clear delivery of bad news, empathic communication,

attentive listening, avoidance of jargon, confirmation of understanding, and discussion of next steps. In prior testing of this instrument, 2 independent raters demonstrated substantial interrater reliability (Cohen's $\kappa = 0.76$).⁴

All checklist scores (pretest compared with posttest within the same study arm) and survey items were compared using the Wilcoxon matched pairs signed rank test. Pretest and posttest checklist scores between study arms (ChatGPT and SP groups) used Wilcoxon rank sum tests. In addition, the pilot data standard deviation was used for a noninferiority power analysis using parameters of alpha 0.05, beta 0.1, and a noninferiority margin of 2 points in checklist scores to determine a sample size for a larger study. Stata 18.0 (Austin, TX) was used to perform all analyses.

RESULTS

Twenty-five trainees (14 anesthesia residents and 11 student registered nurse anesthetists) were enrolled in the study between April and June 2024 with 12 participants randomized to the SP group and 13 randomized to the ChatGPT group. A summary of participants' clinical experience is listed in Table 1. The median years of clinical anesthesia practice was 3 (interquartile range [IQR] 2–7). Twenty-two participants (88%) had experience clinically with difficult conversations. Twelve participants (48%) had previous training in difficult conversations. Participants who received didactic training and subsequently engaged in deliberate practice with a live SP showed an increase from a median pretest 11 (IQR 11–12) to posttest 14 (IQR 14–15; $p = .0005$ (Table 2). Participants who completed didactic instruction and deliberate practice with ChatGPT only also showed improved performance scores from a pretest score of 11.5 (IQR 11–12.5) to a posttest score of 14 (IQR 14–15; $p = .001$). There were no significant differences between the SP or ChatGPT group in pretest ($p = .44$) or posttest ($p = .78$) scores. Both groups also had significant increases in self-reported knowledge, confidence, and ability to apply communication skills (Table 3). Course satisfaction was high for both groups. On average, the SP group practiced 3 scenarios

for 60 minutes, whereas the ChatGPT group practiced an average of 3 scenarios over 29 minutes. Noninferiority testing using an alpha 0.05, beta 0.1 standard deviation of 1, and 2-point margin yielded a total sample size of 14.

DISCUSSION

Our study demonstrates that incorporating didactic training with deliberate practice with ChatGPT can significantly increase anesthesia trainees' simulated communication skills and provided comparable improvements to SP-mediated deliberate practice. Using ChatGPT for deliberate practice in difficult conversations eliminates the need for trained SPs and experienced instructors, allowing learners to practice flexibly on their own schedules without the constraints of staffing availability.⁵ Additionally, ChatGPT can increase learner accessibility because training can be accessed from any computer with an internet connection.⁵ AI enables scalable generation of diverse clinical scenarios, allowing learners to individualize deliberate practice experiences, serving as the cornerstone of Precision Education.⁸

Our study's findings align with other studies that have shown ChatGPT's ability to train physicians and medical students in communication skills.^{5,9,10} We add to these studies' findings with the use of anesthesia trainees, a control group, and incorporation of checklist scoring. Our findings suggest the feasibility of integrating large language models into structured communication curricula within anesthesiology training. Though encouraged by these findings, we recognize that large language model outputs may reflect harmful biases inherent in training data. Although no observable biased responses were encountered during this study, AI-generated content should be vigilantly monitored by content experts when incorporated into educational curricula. In parallel, PHI should not be entered into public-facing large language models without appropriate safeguards. Whereas some institutionally deployed platforms may operate within secure environments (eg, behind local firewalls or via approved privacy gateways), their

continued on next page

continued from previous page

suitability for handling PHI should be determined on a case-by-case basis in consultation with institutional information technology and data security teams.

Several limitations warrant consideration. First, the modalities compared in this study differ fundamentally. The ChatGPT-based training employed in our study was text-based, whereas the SP encounters involved real-time verbal and nonverbal communication. As a result, critical aspects of interpersonal communication, such as eye contact, tone, timing, and body language, were not directly practiced in the ChatGPT group. However, it is important to recognize that effective communication, particularly empathy, can still be conveyed through text-based interactions. Elements of therapeutic communication, such as acknowledgment; validation; and clear, supportive language, can be expressed without reliance on visual or auditory cues. Additionally, all participants underwent pretesting and posttesting using a live SP encounter. This design ensured that communication performance was assessed in a setting requiring integration of both verbal and nonverbal communication skills. As such, improvements observed in the ChatGPT group likely reflect transfer of structured communication behaviors from a text-based training environment to a live interpersonal context.

Another limitation is that both groups received didactic instruction on the SPIKES framework, and this limits our ability to isolate the independent contribution of ChatGPT in observed

improvements. Future larger studies incorporating a didactic-only control arm would help better delineate the relative contributions of didactic instruction and deliberate practice. In addition, use of identical pretest and posttest scenarios may have introduced a testing effect, whereby improved performance was partially attributable to familiarity with the cases rather than the educational intervention alone. Finally, the ChatGPT group had more prior formal training in difficult conversations despite randomization. We were encouraged that, despite having more prior training in the ChatGPT group, pretest scores were similar between groups. Interestingly, the ChatGPT group used significantly less time on deliberate practice; however, both groups completed a similar number of scenarios on average. The shorter duration in the ChatGPT group was likely related to the additional time required to set up scenarios and provide real-time coaching in the in-person format.

As a pilot study, this work is intended to generate preliminary data and highlight the need for larger studies to better isolate the contributions of didactic instruction, deliberate practice, and AI-based tools. Notably, OpenAI has released a voice-enabled chat function capable of simulating conversational tone and emotional nuance, which may provide a more realistic training experience. Future studies incorporating voice-based or multimodal AI interfaces may better replicate the full spectrum of interpersonal communication. Next steps to further this work include using a larger sample size, using ChatGPT for evaluating testing

sessions, and assessing long term retention. In conclusion, AI may prove quite useful to improving physicians' acquisition of nontechnical skills.

References

1. Berkey FJ, Wiedemer JP, Vithalani ND. Delivering bad or life-altering news. *Am Fam Physician*. 2018;98(2):99-104.
2. Boissy A, Windover AK, Bokar D, et al. Communication skills training for physicians improves patient satisfaction. *J Gen Intern Med*. 2016;31(7):755-61.
3. Messerotti A, Banchelli F, Ferrari S, et al. Investigating the association between physicians self-efficacy regarding communication skills and risk of "burnout." *Health Qual Life Outcomes*. 2020;18(1):271.
4. Phillips ML, Tsao M, Davis-Sandfoss A, et al. Use of simulation-based mastery learning curriculum to improve difficult conversation skills among anesthesiologists: a pilot study. *J Educ Perioper Med*. 2023;25(3):E710.
5. Stamer T, Steinhäuser J, Flägel K. Artificial intelligence supporting the training of communication skills in the education of health care professions: scoping review. *J Med Internet Res*. 2023;25:e43311.
6. Baile WF, Buckman R, Lenzi R, et al. SPIKES—a six-step protocol for delivering bad news: application to the patient with cancer. *Oncologist*. 2000;5(4):302-11.
7. Vermeylen JH, Wayne DB, Cohen ER, McGaghie WC, Wood GJ. Promoting readiness for residency: embedding simulation-based mastery learning for breaking bad news into the medicine subinternship. *Acad Med*. 2020;95(7):1050-6.
8. Triola MM, Burk-Rafel J. Precision medical education. *Acad Med*. 2023;98(7):775-81.
9. Webb JJ. Proof of concept: using ChatGPT to teach emergency physicians how to break bad news. *Cureus*. 2023;15(5):e38755.
10. Ji YA, Kim G, Seo JH. Transforming medical communication education: the effectiveness of generative AI education using ChatGPT. *J Health Info Stat*. 2024;49(4):332-8.

continued on next page

continued from previous page

The following authors are at Northwestern University, Feinberg School of Medicine in Chicago, IL: **Meron Amariw** and **Clara S. Miller** are medical students and **Jeffrey H. Barsuk** is a Professor of Medicine and Medical Education and American Board of Internal Medicine Certified in the Department of Medicine and Medical Education. **Mitchell L. Phillips** is an Assistant Professor of Anesthesiology with a Fellowship in Pediatric Anesthesiology and American Board of Anesthesiology (ABA) and Pediatric Anesthesiology Certified in the Department of Pediatric Anesthesiology, Nemours Children's Hospital, Orlando, FL. **John D. Mitchell** is a Professor of Anesthesiology with a Fellowship in Cardiothoracic Anesthesiology, ABA, and National Board of Echocardiography Certified in the Department of Anesthesiology, Henry Ford Health, Detroit, MI. **Heather A. Ballard** is an Associate Professor of Anesthesiology with a Fellowship in Pediatric Anesthesiology and ABA and Pediatric Anesthesiology Certified in the Department of Pediatric Anesthesiology and Medical Education, Ann & Robert H. Lurie Children's Hospital of Chicago, IL.

Note: Meron Amariw and Mitchell L. Phillips contributed equally to the work.

Corresponding author: Mitchell L. Phillips, Department of Pediatric Anesthesiology, Nemours Children's Hospital, 6535 Nemours Parkway, Orlando, FL 32827. Telephone (407) 257-9774.

Email address: Mitchell L. Phillips: Mitchell.Phillips@nemours.org

Note: Clinical trial number and registry URL: ClinicalTrials.gov ID NCT05979090. P.I. Heather Ballard, date: 7-19-2023. Prior Presentations: Society for Education in Anesthesia on April 4, 2025 (Austin, TX), Society for Pediatric Anesthesiology on March 29, 2025 (Orlando, FL).

Funding statement: Support was provided solely from institutional and/or departmental sources.

Abstract

Background: Poor performance during difficult conversations is associated with decreased patient satisfaction and increased clinician burnout. Communication skills can be taught using simulated patients (SPs), but this approach is resource intensive. Artificial intelligence (AI) has emerged as a potential adjunct for communication training. However, its role within anesthesiology education and structured communication curricula remains unclear.

Methods: We conducted a randomized controlled pilot study involving anesthesiology trainees, including anesthesiology residents and student registered nurse anesthetists, at a tertiary pediatric hospital. All participants completed a didactic session on the SPIKES framework, followed by deliberate practice using either OpenAI's ChatGPT or an SP with live instructor feedback. Participants completed pretraining and posttraining assessments using a validated 16-item communication skills checklist with an SP. Self-reported knowledge, confidence, and ability to apply communication skills were assessed using Likert-scale surveys.

Results: Twenty-five trainees were enrolled (ChatGPT n = 13; SP n = 12). Both groups demonstrated significant improvements in checklist scores following training. Self-reported knowledge, confidence, and application of communication skills improved in both groups.

Conclusions: A curriculum combining didactic instruction with deliberate practice using either ChatGPT or SPs was associated with improved simulated communication performance among anesthesiology trainees. This pilot study demonstrates the feasibility of incorporating AI-based tools into communication training.

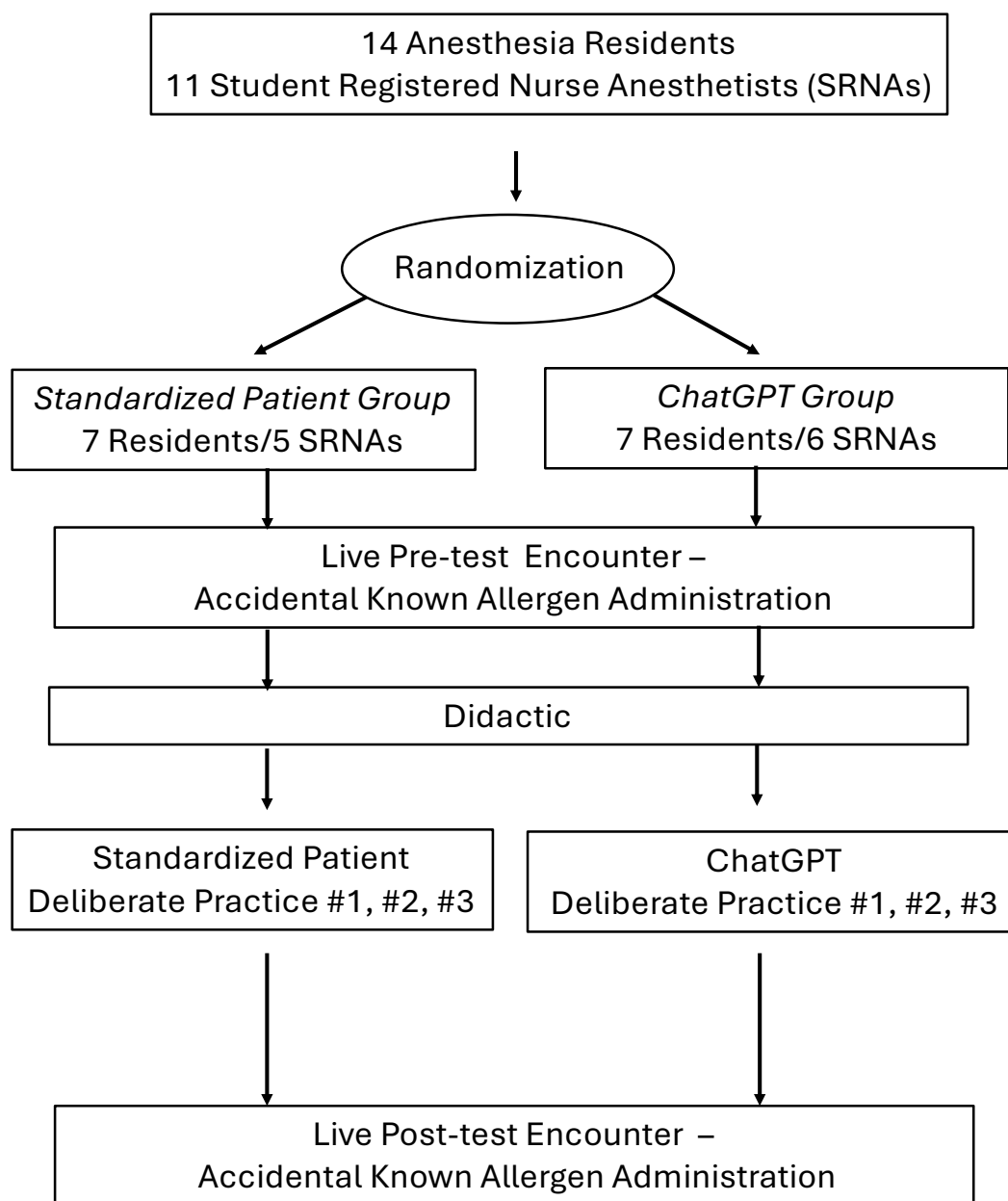
Keywords: Artificial intelligence, communication skills, medical education, anesthesiology, simulation training, deliberate practice

continued on next page

continued from previous page

Figure

Figure 1. Study flow diagram.



Deliberate Practice Scenarios:

#1 – Overdose of Muscle Relaxant

#2 – Accidental Administration of NSAID in Patient with Kidney Disease

#3 – Failed Caudal Block

continued on next page

continued from previous page

Tables

Table 1. Characteristics of Learners Who Participated in Difficult Conversations Curriculum

Group	Standardized Patient (n = 12)	ChatGPT (n = 13)	All (n = 25)
Training, n (%)			
Student nurse anesthetist	5 (41.7)	6 (46.2)	11 (44)
Anesthesia resident	7 (58.3)	7 (53.8)	14 (56)
Clinical experience, y, median (IQR)	4 (2.5–7.5)	3 (2–5)	3 (2–7)
Experience with difficult clinical conversations, n (%)			
Yes	10 (83.3)	12 (92.3)	22 (88)
No	2 (16.7)	1 (7.7)	3 (12)
Number of difficult clinical conversations, n (%)			
Never	2 (18.2)	1 (7.7)	3 (12)
1–5	5 (41.7)	4 (30.8)	9 (36)
6–10	2 (18.2)	4 (30.8)	6 (24)
>10	2 (18.2)	4 (30.8)	6 (24)
Formal Training in Difficult Conversations	4 (33.3%)	8 (61.5)	12 (48)
When was training in difficult conversations? n (%)			
Never	8 (66.7)	5 (38.5)	13 (52)
Medical school	4 (33.3)	3 (23.1)	7 (28)
Nurse anesthetist school	0	1 (7.7)	1 (4)
Residency	0	4 (30.8)	4 (16)

Table 2. Checklist Test Scores Between Standardized Patient and ChatGPT Deliberate Practice

Group	Standardized patient (n = 12)	ChatGPT (n = 13)	P-value
Pretest score (0–16), median (IQR)	11 (11–12)	11.5 (11–12.5)	.445
Posttest score (0–16), median (IQR)	14 (14–15)	14 (14–15)	.782
p-value between pretest and posttest	.0005	.001	

continued on next page

continued from previous page

Tables

Table 3. Self-Reported Knowledge, Confidence, Application and Course Satisfaction Comparisons Between Standardized Patient and ChatGPT Deliberate Practice

Group	Standardized patient (n = 12)	ChatGPT (n = 13)	P-value
Knowledge before training, median (IQR) (1: little knowledge to 5: superior knowledge)	2.5 (2–3)	3 (2–4)	.374
Knowledge after training, median (IQR) (1: little knowledge to 5: superior knowledge)	5 (4–5)	4 (4–5)	.173
p-value between before and after training	.0005	.003	
Confidence before training, median (IQR) (1: little confidence to 5: superior confidence)	2.5 (2–3)	3 (2–4)	.783
Confidence after training, median (IQR) (1: little confidence to 5: superior confidence)	4 (4–4)	4 (4–5)	.407
p-value between before and after training	.0005	.0002	
Application before training, median (IQR) (1: little application to 5: superior application)	2.5 (1.5–4)	3 (2–4)	.791
Application after training, median (IQR) (1: little application to 5: superior application)	5 (4.5–5)	5 (4–5)	.998
p-value between before and after training	.0005	.0002	
Course satisfaction, median (IQR) (1: not satisfied to 5: very satisfied)	5 (4–5)	5 (5–5)	.424

continued on next page

continued from previous page

Supplemental Online Materials

Supplemental 1: ChatGPT Prompt and Practice Scenario Example

I am an anesthesia practitioner. I would like to do deliberate practice on braking bad news to a parent after surgery. I would like you to be my teacher. You will simulate a detailed scenario, which I will provide below, in which I will have to engage in a difficult dialogue with a parent. You will first frame the scenario. Then you will begin roleplaying as the parent. You will only respond in first-person as a parent would respond. You will wait for my response in each step of the scenario. Let's go through several rounds of responses until the case concludes. After you conclude the case, please grade my response using the SPIKES and NURSES framework for breaking bad news.

Tyler is a 7-month-old, 7 kg male who presents for bilateral laparoscopic inguinal herniorrhaphy under general anesthesia. He has no other significant medical history besides bilateral inguinal hernia. He was born full term and went home shortly after birth.

He underwent an uncomplicated anesthetic. The surgeon is almost done with the surgery, but comments that the patient seems inadequately relaxed. In a haste to administer muscle relaxant, you accidentally administer 40 mg of rocuronium.

The surgery is now finished and the patient has no twitches. Unfortunately, the location that you are practicing at does not have sugammadex. The surgeon, who has a full schedule, would like the patient taken to the ICU where he can wait until his muscle relaxant wears off.

The surgeon asks you to discuss the medication error with the family.

continued on next page

continued from previous page

Supplemental Online Materials

Supplemental 2: Breaking Bad News Skill Checklist

	Yes	No
Creates initial rapport when first walking into room (e.g. introduces self/introduced by someone)		
Sits down		
Assumes a comfortable interpersonal distance		
Assesses family's perception or understanding of medical situation before breaking news (e.g. "tell me what you understand")		
Asks permission before giving the news (e.g. "I would like to discuss what happened")		
Gives a clear and concise "warning shot" (e.g. "I have some serious news")		
Pauses after delivering bad news		
Delivers bad news within the first minute of the conversation		
Delivers an empathic statement (e.g. "I know this is not what you expected to hear today")		
Suggests a plan for the next step		
Ensures family understanding (e.g. "It sounds like ...")		
Avoids medical jargon (uses technical language without clarifying what it means)		
Gives information in small chunks (e.g. no more than 1 chunk of information before allowing family to process)		
Avoids giving information while family very emotional		
Avoids providing reassurances to family's emotion (e.g. avoid saying something like "it's ok")		
Listens attentively		