Application of an Artificial Intelligence Simulated Patient Model in Teaching Neurosurgical Decision-Making

Ann. Ital. Chir., 2026 97, 11: 1971–1979
https://doi.org/10.62713/aic.4849

Sheng Ruan[1], Shanping Ruan[1], Jianfeng Wang[1], Xiaoming Hu[1], Ting Guo[1], Jinping Zhou[2]

[1]Department of Neurosurgery, Taizhou Hospital of Zhejiang Province, 317000 Taizhou, Zhejiang, China
[2]Department of Emergency, Taizhou Hospital of Zhejiang Province, 317000 Taizhou, Zhejiang, China

AIM: Training in neurosurgical decision-making remains a challenging component of standardized residency training. Traditional teaching models are limited in their ability to simulate the uncertainty and complexity encountered in real-world clinical scenarios. As a novel educational tool, the potential value of artificial intelligence (AI)-simulated patient models in surgical decision-making training requires further evaluation. This study aimed to explore the application of an AI-simulated patient model in neurosurgical decision-making education and to assess its effectiveness in improving residents' clinical decision-making abilities.

METHODS: A retrospective study was conducted involving 157 residents who underwent standardized residency training in the Department of Neurosurgery from January 2023 to April 2026. Based on the implementation of the AI-simulated teaching model beginning in January 2025, participants were divided into two groups according to the teaching model used: an experimental group (n = 72, receiving AI-simulated patient teaching from January 2025 to April 2026) and a control group (n = 85, receiving traditional case discussion teaching from January 2023 to December 2024). The experimental group used a large language model-based virtual patient system powered by DeepSeek to perform dialogue-driven history taking, image interpretation, surgical planning, and risk prediction with real-time feedback, while the control group received conventional methods. After teaching, an objective structured clinical examination (OSCE) was used to assess surgical decision-making across five dimensions, including localization diagnosis, surgical approach, risk identification, emergency management, and surgical timing, and a questionnaire was used to measure teaching satisfaction.

RESULTS: The total score for surgical decision-making ability was significantly higher in the experimental group than in the control group ($p < 0.001$). Dimensional analysis showed that localization diagnosis, surgical method selection, risk identification, emergency response, and surgical timing judgment were higher in the experimental group than in the control group (all $p < 0.05$). The excellent rate for each assessment dimension (>85 points) was significantly higher in the experimental group than in the control group ($p < 0.05$). The satisfaction score was also higher in the experimental group than in the control group ($p < 0.001$).

CONCLUSIONS: The AI-simulated patient model based on a general-purpose large language model was associated with better surgical decision-making performance among neurosurgery residents. It has significant advantages in localization diagnosis, surgical approach selection, risk identification, emergency response, and judgment of surgical timing, and is well accepted by residents.

Keywords: artificial intelligence; neurosurgery; surgical decision making; virtual patient simulation; medical education

Introduction

Neurosurgery is characterized by complex anatomical structure, high operative risks, and a low tolerance for errors in decision-making. The development of surgical decision-making ability is directly related to patient safety and prognosis [1]. One of the core goals of standardized residency training in neurosurgery is to cultivate residents' ability to make surgical decisions independently, accurately, and efficiently in complex clinical situations. However, the development of this ability requires the long-term accumulation of clinical experience and systematic cognitive training. Achieving this goal within the limited period of standardized training has become a major focus and challenge in neurosurgical education reform.

Traditional surgical decision-making training mainly relies on case discussions and clinical teaching. The case discussion model helps residents establish a basic framework for clinical thinking through the collective analysis of typical cases and teachers' comments. In clinical teaching, residents can observe and learn the decision-making process in actual surgical practice through a mentorship system. However, both models have obvious limitations. On the one hand, owing to limited clinical resources and medical safety requirements, it is difficult for residents to independently complete the entire decision-making process in real clinical settings, and even more difficult for them to receive imme-

Submitted: 16 June 2026 Revised: 29 July 2026 Accepted: 7 August 2026 Published: 10 November 2026
Correspondence to: Jinping Zhou, Department of Emergency, Taizhou Hospital of Zhejiang Province, 317000 Taizhou, Zhejiang, China (e-mail: zhoujp7470@enzemed.com).
Editor: Abuzer Güngör

diate feedback and opportunities for correction after making incorrect decisions [2]. On the other hand, case discussions lack dynamic interaction, making it difficult to reproduce the uncertainty, time pressure, and crisis situations commonly encountered in clinical decision-making. Although problem-based learning (PBL) can cultivate residents' dialectical thinking ability to a certain extent, its simulated cases also lack real-time interaction [3]. Therefore, exploring surgical decision-making training models that provide safe, repeatable, and dynamic interaction has important educational value and clinical significance [4].

In recent years, the application of artificial intelligence (AI) technology in the field of medical education has made significant progress [5,6]. Generative AI has shown increasing potential in medical education, including clinical reasoning training and virtual patient interaction. Among them, AI-simulated patient models based on large language models (LLMs) have become an active area of research. Previous studies have demonstrated the feasibility of AI-simulated patients in clinical skills training and the potential benefits of combining AI with human instruction in surgical education [7,8]. In neurosurgical education, LLM-based tools have also been explored for perioperative patient communication [9]. However, existing research has mainly focused on basic clinical skills, such as history taking and physical examination, with relatively limited attention to higher-order cognitive abilities such as surgical decision-making. Given the complexity, uncertainty, and time-sensitive nature of neurosurgical decisions, interactive teaching approaches are needed to help residents integrate clinical information, select appropriate surgical strategies, and identify potential risks.

Despite these developments, high-quality evidence regarding the application of AI-simulated patient models in neurosurgical decision-making education remains limited. Based on the above background, this study aimed to explore the application of an AI-simulated patient model based on a large language model (DeepSeek, 2025 version, DeepSeek Inc., https://chat.deepseek.com) in neurosurgical decision-making teaching, with the aim of providing an evidence-based basis for optimizing standardized residency training. This study focused on five core dimensions of decision-making: localization diagnosis, surgical approach selection, risk identification, emergency response, and judgment of surgical timing. A retrospective study was conducted to compare post-teaching surgical decision-making performance between the AI-simulated patient model and the traditional case discussion teaching model and to provide a reference for the reform of neurosurgical clinical education.

Methods

Participants

A total of 157 residents who underwent standardized training in the Department of Neurosurgery at Taizhou Hospital of Zhejiang Province from January 2023 to April 2026 were retrospectively selected as the study participants. Among them, 85 residents included between January 2023 and December 2024 received traditional case discussion teaching (control group), and 72 residents included between January 2025 and April 2026 received AI-simulated patient teaching (experimental group). Inclusion criteria: (1) completion of the neurosurgical rotation plan; (2) no previous neurosurgical training experience; (3) complete clinical data. Exclusion criteria: (1) withdrawal from rotation; (2) failure to complete all teaching and assessment components. This study was a retrospective educational study, using routine teaching assessment data, without the use of patient information. All case scripts used for AI teaching were derived from de-identified clinical cases, with all patient identifiers (names, medical record numbers, contact information, and other protected health information) removed prior to use. Residents' interaction data with the AI system were not recorded or stored because of limitations of the platform interface; no personally identifiable information was collected or transmitted to the AI platform. The study protocol, including the use of AI-assisted teaching methods for residency education and the retrospective analysis of anonymized teaching assessment data, was reviewed and approved by the ethics committee of Taizhou Hospital of Zhejiang Province (approval number: K20260602) and was conducted in accordance with the Declaration of Helsinki.

Teaching Methods

All teaching sessions for both groups were conducted by the same group of clinical instructors. Prior to the initiation of the study, all instructors participated in a standardization training session to ensure consistent understanding of the teaching objectives, case content, and assessment criteria across both teaching models. The same standardized case scripts were used for both groups to ensure consistency of teaching content.

Experimental Group: AI-Simulated Patient Teaching Model Based on a Large Language Model

The experimental group adopted the AI-simulated patient teaching model based on a large language model. In this model, a general-purpose large language model (DeepSeek, https://chat.deepseek.com) was used to build a virtual patient interaction system, without the need to develop a special software platform.

Case Script Design. Senior neurosurgeons (associate chief physician or above) in the research group wrote standardized case scripts according to actual clinical cases. A total of 8 core cases were designed, covering common disease types in neurosurgery: intracranial aneurysm (2 cases), glioma (2 cases), pituitary tumor (1 case), acoustic neuroma (1 case), chronic subdural hematoma (1 case), and hypertensive intracerebral hemorrhage (HICH) (1 case). All eight core case scripts and their corresponding variant cases

were developed using a standardized template, with predefined decision nodes and evidence-based feedback criteria to ensure consistency across teaching sessions. Each case script contained the following elements: medical history: chief complaint, present history, past history, family history; physical examination: key points of nervous system examination; Imaging description: text description of key imaging features such as head computed tomography (CT), computed tomography angiography (CTA), and magnetic resonance imaging (MRI) (imaging information was presented in text form due to platform limitations); Key decision nodes: five core decision points: localization diagnosis, surgical selection, risk identification, emergency response, and surgical timing judgment; Preset feedback basis: evidence-based information and risk warnings corresponding to each decision node.

Teaching Implementation Process. The AI-simulated patient was used as a "virtual patient + virtual teacher" to participate in the teaching process. The specific procedures were as follows: ① Initialization: Before teaching, the teacher set the AI tool as a simulated patient and input standardized prompts: "You are a simulated patient of a neurosurgery resident; please answer questions according to the following case script. You will provide medical history in the voice of a real patient, but will not voluntarily disclose the diagnosis. Please provide relevant information about the patient's condition one by one in response to their questions." ② Inquiry and physical examination stage: the trainees interacted with the AI-simulated patients as the attending physician, inquired about their medical history, and independently obtained the physical examination findings. According to the preset case script, the AI responded to the residents' questions one by one using natural language to simulate the real interrogation process. Trainees were expected to collect complete clinical information through systematic history taking. ③ Image interpretation stage: the teacher provided real imaging data (Digital Imaging and Communications in Medicine (DICOM) format or image pictures) corresponding to the cases to the residents, and the residents were required to independently interpret the imaging findings and make a localization diagnosis in combination with the information obtained during history taking. The AI simulated patient did not directly participate in image interpretation at this stage but provided analytical feedback on the imaging findings after the trainees had completed their diagnosis. ④ Surgical decision-making stage: after completing the diagnosis, the trainees explained the surgical plan to the AI-simulated patient (switched to the role of "virtual teacher" at this time), including surgical timing selection, surgical approach, intraoperative risk prediction, and emergency plan formulation. According to the database of predetermined decision nodes, the AI provided evidence-based feedback, including evidence supporting the selected surgical time, the indications and contraindications for surgical procedures, potential risks and preventive measures, and differences in prognosis among different decision pathways. ⑤ Real-time feedback and summary: after the teaching of each case, the AI generated a structured feedback report that included the logical completeness of the trainee's inquiry, the adequacy of the diagnostic basis, the accuracy of surgical timing judgment, the rationality of surgical procedure selection, and the comprehensiveness of risk identification. The teachers provided supplementary comments based on the AI feedback to strengthen the trainees' understanding of the key decision points.

Teaching Arrangement. AI-simulated patient teaching was conducted once a week, with each session lasting 90 minutes, for a total of 8 sessions. Each teaching session consisted of one main case and one variant case (involving a different clinical type or complication of the same disease).

Quality Control. In order to ensure the stability and accuracy of the outputs generated by the AI-simulated patient model, the following measures were taken: before each teaching session, the teacher initialized the AI system and entered standardized role-setting instructions to minimize output variability; the standard answers and feedback basis of key decision nodes were preset, and the consistency of the AI output was reviewed by the teachers. The teacher was present throughout the teaching process, and manual intervention and correction were performed when the AI-generated output deviated from the preset case. Eight rounds of instruction for the same case used a standardized cue word template to ensure that all trainees received the same initial conditions.

Technical Specifications of the AI System. The large language model used in this study was DeepSeek (2025 version, DeepSeek Inc.), a general-purpose large language model accessed via its web interface (https://chat.deepseek.com). The complete case script template included the following elements: chief complaint, present history, past history, physical examination findings, imaging description, and five key decision nodes (localization diagnosis, surgical approach selection, risk identification, emergency response, and surgical timing judgment), each with a preset evidence-based feedback basis. A representative case example (hypertensive intracerebral hemorrhage) is detailed in **Supplementary Text 1** to illustrate the template structure. Due to the platform interface limitations, the AI dialogue logs were not systematically stored. The standardized prompt template used for the AI initialization was as follows: "You are a simulated patient in a neurosurgery residency training scenario. You will adopt the persona of a real patient and provide medical history and physical examination findings based on the following case script. Do not voluntarily disclose the diagnosis. Respond to the resident's

questions one by one in natural language. When the resident completes the diagnosis and proposes a surgical plan, switch to the role of a virtual teacher and provide evidence-based feedback on the decision.".

Control Group: Traditional Case Discussion Teaching

The control group adopted the traditional case discussion teaching model. The clinical teachers selected the same neurosurgical cases as those used in the experimental group, presented the case data in PowerPoint format, organized the students into groups for discussion, and finally summarized and commented on the cases. The teaching duration and frequency in the control group were consistent with those in the experimental group.

Effect Evaluation

Assessment of Surgical Decision-Making Ability

The objective structured clinical examination (OSCE) was used to evaluate the surgical decision-making ability of the two groups after the teaching sessions. Four standardized cases (intracranial aneurysm, glioma, hypertensive intracerebral hemorrhage, pituitary tumor) were set up for the assessment, and each case included five stations: localization diagnosis, surgical approach selection, risk identification, emergency treatment, and surgical timing judgment. Each station was scored on a 100-point scale. A structured marking scheme was used for each station, with scores allocated based on predefined key decision-making elements (e.g., accuracy of diagnosis, appropriateness of surgical approach, comprehensiveness of risk identification, adequacy of the emergency management plan, and accuracy of surgical timing judgment). A passing score of 60 was set for each station. For each case, the case score was calculated as the average of its 5 station scores. For each of the five dimensions (e.g., localization diagnosis), the dimension score for each resident was calculated as the average of the corresponding station scores across the four cases. The total examination score was then calculated as the average of the five-dimension scores (equivalently, the average of the four case scores), yielding a single overall score ranging from 0 to 100 for each resident. An "excellent" rating was defined as a dimension score strictly greater than 85, and the excellent rate for each dimension was calculated as the proportion of residents scoring >85 on that dimension. The assessment used double-blind scoring and was independently scored by two senior neurosurgeons with the rank of deputy chief physicians; the average value of the two scores was used as the final score. Prior to the formal assessment, all examiners participated in a standardized training session to ensure consistent understanding and application of the scoring criteria across all stations.

Teaching Satisfaction Evaluation

A self-designed questionnaire was used to assess student satisfaction with the teaching model, including four items: teaching format, practicality of the content, perceived improvement in ability, and overall evaluation. A 5-point Likert scale was used (1 = very dissatisfied, 5 = very satisfied). The internal consistency of the questionnaire was assessed using Cronbach's α, which was 0.822, indicating good internal consistency.

Statistical Methods

SPSS 26.0 software (IBM Corp., Armonk, NY, USA) was used for statistical analysis. Measurement data were tested for normality, and data conforming to a normal distribution were expressed as mean ± standard deviation ($\bar{x} \pm s$). An independent- sample *t*-test was used for comparison between groups. Non-normally distributed data were expressed as median (interquartile range) [M (Q_1, Q_3)], and the Mann-Whitney U test was used for comparison between groups. The count data were expressed as frequency (%), and the χ^2 test was used for comparison between groups. $p < 0.05$ was considered statistically significant. Effect sizes were calculated for all significant between-group comparisons. For the Mann-Whitney U test, the rank-biserial correlation (r) was calculated as $r = Z / \sqrt{N}$. For the χ^2 test, Cramér's V was calculated as $V = \sqrt{(\chi^2/N)}$.

Results

Baseline Comparison

A total of 157 residents were enrolled, including 72 in the experimental group and 85 in the control group. There were no significant differences in the examination scores, age, gender composition, and educational background between the two groups ($p > 0.05$) (Table 1).

Comparison of Surgical Decision-Making Ability

The median score of surgical decision-making ability was 82.97 in the experimental group and 76.45 in the control group, and the difference was statistically significant ($p < 0.001$). The analysis of each dimension showed that the scores of the five dimensions were higher in the experimental group than in the control group, including localization diagnosis, surgical approach selection, risk identification, emergency response, and surgical timing judgment, and the differences were statistically significant (all $p < 0.01$) (Table 2).

Comparison of Scores Across Different Case Types

In order to evaluate the effect of the AI-simulated patient model on decision-making ability across different disease types, the four OSCE cases were categorized as aneurysm, glioma, cerebral hemorrhage, and pituitary tumor according to disease type, and the mean scores of the two groups for each case type were compared. The results showed that the case-type scores for all four disease categories were higher in the experimental group than in the control group, and the differences were statistically significant ($p < 0.001$). The lowest scores were observed in the case of pituitary ade-

Table 1. Comparison of baseline data between the two groups of trainees.

Variables	Total (n = 157)	Experimental group (n = 72)	Control group (n = 85)	Statistic	*p*
Department-entry examination scores, M (Q_1, Q_3)	72.00 (70.00, 78.00)	72.50 (70.00, 78.25)	72.00 (69.00, 75.00)	Z = –1.23	0.217
Age, M (Q_1, Q_3)	25.00 (24.00, 26.00)	25.50 (24.75, 26.25)	25.00 (24.00, 26.00)	Z = –0.14	0.885
Gender, n (%)				χ^2 = 0.01	0.941
Male	79 (50.32)	36 (50.00)	43 (50.59)		
Female	78 (49.68)	36 (50.00)	42 (49.41)		
Education background, n (%)				χ^2 = 0.06	0.809
Bachelor	117 (74.52)	53 (73.61)	64 (75.29)		
Master	40 (25.48)	19 (26.39)	21 (24.71)		

Table 2. Comparison of surgical decision-making ability scores between the two groups of trainees.

Variables	Total (n = 157)	Experimental group (n = 72)	Control group (n = 85)	Statistic	*p*	Effect size
Localization diagnosis, M (Q_1, Q_3)	83.00 (78.25, 86.00)	85.00 (83.00, 87.00)	79.25 (75.25, 83.25)	Z = –6.60	<0.001	0.53
Selection of surgical approaches, M (Q_1, Q_3)	80.25 (74.25, 83.25)	83.00 (81.00, 85.00)	75.25 (71.25, 79.25)	Z = –7.76	<0.001	0.62
Identification of risks, M (Q_1, Q_3)	79.75 (73.75, 82.75)	81.75 (79.75, 83.75)	74.75 (70.75, 78.75)	Z = –7.41	<0.001	0.59
Emergency response, M (Q_1, Q_3)	81.25 (75.25, 84.50)	83.50 (81.50, 85.50)	76.25 (72.25, 80.25)	Z = –7.47	<0.001	0.60
Timing of Surgery, M (Q_1, Q_3)	78.75 (76.00, 84.25)	80.75 (77.19, 85.75)	77.25 (75.75, 81.75)	Z = –2.76	0.006	0.22
Total examination score, M (Q_1, Q_3)	80.45 (75.75, 83.85)	82.97 (80.60, 84.91)	76.45 (73.30, 81.10)	Z = –6.82	<0.001	0.54

Table 3. Comparison of examination scores of different case types between the two groups.

Variables	Total (n = 157)	Experimental group (n = 72)	Control group (n = 85)	Statistic	*p*	Effect size (r)
Intracranial aneurysms, M (Q_1, Q_3)	86.40 (83.40, 89.40)	87.40 (85.40, 89.40)	84.80 (80.80, 88.80)	Z = –3.63	<0.001	0.29
Brain glioma, M (Q_1, Q_3)	79.40 (73.00, 82.40)	82.40 (80.40, 84.40)	74.00 (70.00, 78.00)	Z = –8.45	<0.001	0.67
Hypertensive intracerebral hemorrhage, M (Q_1, Q_3)	79.00 (72.00, 82.00)	82.00 (80.00, 84.00)	73.00 (69.00, 77.00)	Z = –8.76	<0.001	0.70
Pituitary adenoma, M (Q_1, Q_3)	78.20 (71.80, 81.20)	81.20 (79.20, 83.20)	72.80 (68.80, 76.80)	Z = –8.45	<0.001	0.67

noma, while the smallest effect size was observed for intracranial aneurysms (r = 0.29) (Table 3).

Comparison of Excellent Rates Across Assessment Dimensions

To further analyze the teaching effects of the two teaching models across different ability dimensions, the proportions of trainees who achieved excellent scores (>85 points) in each assessment dimension were compared between the two groups. The results showed that the excellent rates for all five dimensions were higher in the experimental group than in the control group, and the differences were statistically significant ($p < 0.05$). Among them, the differences in the excellent rates of surgical selection and risk identification were relatively more pronounced ($p < 0.05$), suggesting that the training effect of the AI-simulated patient model on these abilities was particularly prominent (Table 4).

Comparison of Teaching Satisfaction Between the Two Groups

The median score of teaching satisfaction was 17.00 (16.00, 19.25) in the experimental group and 16.00 (14.00, 17.00) in the control group, and the difference was statistically significant (Z = –5.09, $p < 0.001$). Analysis of each dimension showed that satisfaction in the experimental group was higher than in the control group across all four dimensions: teaching format, practicality of the content, perceived improvement in ability, and overall evaluation, and the differences were statistically significant ($p < 0.01$) (Table 5).

Discussion

In this study, the teaching data of 157 neurosurgery residents were retrospectively analyzed, and the application of an AI-simulated patient model based on a large language model (DeepSeek, https://chat.deepseek.com) in surgical decision-making teaching was systematically evaluated. The results showed that the experimental group achieved significantly higher scores than the control group in five dimensions: localization diagnosis, surgical approach selection, risk identification, emergency response, and judgment of surgical timing. The excellent rates for all five dimensions were also higher, and residents' teaching satisfaction was greater in the experimental group. These

Table 4. Comparison of the excellent rates of each assessment between the two groups.

Variables	Total (n = 157)	Experimental group (n = 72)	Control group (n = 85)	Statistic	*p*	Effect size
Localization diagnosis, n (%)				$\chi^2 = 8.73$	0.003	0.24
No	110 (70.06)	42 (58.33)	68 (80.00)			
Yes	47 (29.94)	30 (41.67)	17 (20.00)			
Selection of surgical approaches, n (%)				$\chi^2 = 12.02$	0.001	0.28
No	136 (86.62)	55 (76.39)	81 (95.29)			
Yes	21 (13.38)	17 (23.61)	4 (4.71)			
Identification of risks, n (%)				$\chi^2 = 11.12$	<0.001	0.27
No	142 (90.45)	59 (81.94)	83 (97.65)			
Yes	15 (9.55)	13 (18.06)	2 (2.35)			
Emergency response, n (%)				$\chi^2 = 5.42$	0.020	0.19
No	126 (80.25)	52 (72.22)	74 (87.06)			
Yes	31 (19.75)	20 (27.78)	11 (12.94)			
Timing of Surgery, n (%)				$\chi^2 = 8.48$	0.004	0.23
No	125 (79.62)	50 (69.44)	75 (88.24)			
Yes	32 (20.38)	22 (30.56)	10 (11.76)			

Table 5. Comparison of teaching satisfaction scores between the two groups.

Variables	Total (n = 157)	Experimental group (n = 72)	Control group (n = 85)	Statistic	*p*	Effect size
Total score, M (Q_1, Q_3)	16.00 (15.00, 18.00)	17.00 (16.00, 19.25)	16.00 (14.00, 17.00)	Z = –5.09	<0.001	0.41
Teaching format, M (Q_1, Q_3)	4.00 (3.00, 5.00)	4.50 (4.00, 5.00)	4.00 (3.00, 4.00)	Z = –4.16	<0.001	0.33
Practicality of the content, M (Q_1, Q_3)	4.00 (4.00, 4.00)	4.00 (4.00, 5.00)	4.00 (4.00, 4.00)	Z = –3.19	0.001	0.25
Perceived improvement in ability, M (Q_1, Q_3)	4.00 (4.00, 5.00)	4.00 (4.00, 5.00)	4.00 (3.00, 4.00)	Z = –4.29	<0.001	0.34
Overall evaluation, M (Q_1, Q_3)	4.00 (4.00, 5.00)	4.50 (4.00, 5.00)	4.00 (3.00, 4.00)	Z = –5.58	<0.001	0.45

findings provide new evidence supporting the application of AI-assisted teaching in neurosurgery and suggest the potential application of a lightweight AI-simulated patient model in teaching higher-order clinical decision-making skills.

Training in neurosurgical decision-making has always been a challenging aspect of standardized residency training. Traditional teaching models have difficulty simulating the dynamics and uncertainty of real-world clinical situations. The results of this study showed that the median total score for surgical decision-making ability was significantly higher in the experimental group than in the control group, with a highly statistically significant difference ($p < 0.001$). This finding is consistent with the results of the randomized controlled trial conducted by McCarrick et al. [7], in which ChatGPT was used as a simulated patient for surgical history-taking training, and the intervention group showed a significant improvement in OSCE scores. However, this study further extended the application of the AI-simulated patient model from history taking to the higher-order cognitive domain of surgical decision-making, with a larger sample size (157 residents) and more comprehensive teaching dimensions. In addition, some studies have found that AI-enhanced human-machine combined teaching is significantly more effective than AI-only teaching in improving surgical performance, particularly in the control of bleeding risk and normal tissue damage [8]. The results of the present study further support this conclusion and extend its application to neurosurgical decision-making education. The greater improvement observed in the experimental group may be associated with the immediate feedback mechanism provided by the AI-simulated patient model. The system can immediately indicate whether the basis for a decision is appropriate. This "trial-and-error-feedback-correction" learning model may help accelerate the development of clinical reasoning ability, thereby achieving greater improvement in decision-making ability within a relatively short period.

Handling emergencies in neurosurgery is an important component of clinical decision-making; however, students rarely have the opportunity to practice such scenarios in traditional teaching settings. The AI-simulated patient model can simulate the occurrence and development of critical events and train students' ability to make rapid decisions under pressure. Previous studies have shown that simulation-based training has unique value in cultivating non-technical skills related to surgery. Compared with tra-

ditional case discussion and clinical teaching, simulation-based teaching can provide students with a safe and repeatable crisis scenario training environment, allowing learners to repeatedly practice the emergency response process without medical risk and receive immediate feedback during the simulation process. This training mode is especially suitable for neurosurgery and other high-risk specialties, as it can effectively compensate for the limited opportunities students have to encounter real critical cases in traditional teaching [10,11]. The results of this study further suggested that AI-simulated teaching may help improve residents' emergency response-related decision-making skills, as reflected by their OSCE performance. This approach provides an interactive learning environment in which residents can practice emergency management decisions and receive feedback, although its effectiveness in improving real-world emergency response under time pressure and uncertainty requires further investigation. This finding provides new empirical support for exploring a safe and efficient surgical decision-making training model and suggests a potential direction for integrating AI-simulated technology into the standardized training system of surgical residents in the future.

The experimental group also performed well in the dimension of surgical timing judgment. The surgical timing for neurosurgical diseases is a key factor affecting prognosis: treatment within 72 hours after aneurysm rupture can significantly reduce the risk of rebleeding [12]. Ultra-early surgery for hypertensive intracerebral hemorrhage (HICH) (within 24 hours of onset) helps improve functional outcomes [13,14]. In traditional teaching, the timing of surgery is often taught as a knowledge point, and students cannot determine the appropriate timing in response to dynamic changes in disease status. AI-simulated patients can simulate the temporal evolution of disease (such as the progression of cerebral vasospasm following subarachnoid hemorrhage and the peak of edema after intracerebral hemorrhage), thereby allowing students to experience the key decision point of "when to intervene" in the simulation situation, which may contribute to strengthening their time-sensitive decision-making ability.

This study further analyzed the teaching effects of an AI-simulated patient model across different disease types. The results showed that the test scores for intracranial aneurysm, glioma, hypertensive intracerebral hemorrhage, and pituitary tumor were significantly better in the experimental group than in the control group. This finding has important implications for clinical teaching, indicating that the AI-simulated patient model may have broad applicability in training decision-making ability across different types of neurosurgical diseases. Hussain et al. [15] demonstrated that both ChatGPT and DeepSeek showed good applicability in generating patient education guidelines for stroke imaging, suggesting the potential utility of general-purpose AI models in specific neurosurgical education contexts. The results of this study suggest that the AI-simulated patient system based on DeepSeek may be capable of supporting surgical decision-making education for a variety of neurosurgical diseases and may facilitate the centralized development of clinical teaching resources.

In terms of teaching satisfaction, the evaluation of the teaching model in the experimental group was significantly higher than in the control group, particularly in the dimensions of teaching format and overall evaluation. The proportions of "5-point" ratings in the experimental group were numerically higher than those in the control group. This result is consistent with the findings of Zeng et al. [6] who pointed out that virtual patients driven by large language models are highly interactive and personalized, which can improve students' learning engagement. The reason for this may be that the AI-simulated patient model offers the advantages of relatively low development cost and rapid deployment, and students can complete learning through natural language dialogue, thereby reducing the technical barriers to use. Meanwhile, the immediate feedback mechanism enables students to clearly identify their own knowledge gaps and decision-making deviations, thereby enhancing the pertinence and effectiveness of learning.

The innovation of this study is mainly reflected in the following aspects: first, innovation in the research paradigm. Most previous studies on AI patient simulation focused on the training of basic skills such as history taking. This study applied the AI-simulated patient model to the higher-order cognitive domain of surgical decision-making, thereby expanding the application scope of AI in neurosurgical education. Second, innovation in the evaluation dimensions. This study constructed an evaluation system covering five dimensions, including localization diagnosis, surgical approach selection, risk identification, emergency response, and surgical timing judgment, which can more precisely reflect the effects of AI teaching on the improvement of different decision-making abilities. Third, innovation in the teaching scenario. In this study, the AI-simulated patient was designed with the dual identity of "virtual patient + virtual teacher", and a closed-loop learning model of "practice-feedback-correction" was realized. It is important to emphasize that the AI-simulated patient model is designed as an educational aid to complement, rather than replace, traditional clinical training. The supervising teacher played an essential role in interpreting AI feedback, correcting inaccuracies, and providing contextualized clinical reasoning that the current AI system cannot offer.

Our study has several limitations. First, this study was a single-center retrospective study with limited sample representativeness. In addition, due to the retrospective nature of this study, certain baseline variables that may influence OSCE performance—including postgraduate year, prior neurosurgical exposure, prior AI use, and rotation duration—were not systematically documented and therefore could not be included in the baseline comparison. Al-

though the available baseline data (department entry examination scores, age, gender, and educational background) were comparable between the two groups, the lack of these additional variables may have introduced residual confounding. Furthermore, although the OSCE scoring was performed independently by two senior physicians, inter-rater reliability was not calculated, which may have introduced scorer variability. Further multicenter prospective studies are needed to verify the generalizability of these conclusions. Second, the evaluation indicators are mainly short-term teaching effects, and there is a lack of long-term tracking of students' clinical practice performance. Whether the observed improvements in OSCE performance translate into sustained competency in real-world clinical settings remains unknown. In addition, the novelty effect of the AI tool—whereby residents' engagement and performance may have been temporarily boosted by the mere introduction of a new technology—cannot be excluded. Similarly, residents' inherent enthusiasm for or interest in AI technology may have contributed to the observed differences, independent of the pedagogical efficacy of the AI system itself. Future research should combine intraoperative performance evaluation and clinical outcome data to evaluate the practical value of AI simulation teaching more comprehensively. Third, there are still differences between the AI simulation system and real clinical settings. Whether the decision-making ability of trainees shown in the simulation environment can be fully transferred to clinical practice needs to be further verified. Fourth, this study used a retrospective non-randomized design, and the two groups of students were enrolled in stages. The AI-simulated patient teaching model was introduced as a new educational intervention in January 2025 within the existing standardized residency training program. Given the implementation of teaching models within the established training schedule, individual random allocation to different teaching models was not feasible; therefore, historical cohorts before and after the implementation of AI-assisted teaching were compared. Although the baseline data were comparable, potential time-period bias cannot be completely excluded. Factors that may have changed over time include the accumulation of teaching experience among clinical instructors, the optimization of teaching content and case materials, the refinement of assessment standards, adjustments to the residency rotation curriculum structure, and the increasing institutional familiarity with AI technology. These temporal changes may affect the internal validity of the comparison between the groups. In addition, although the teaching duration and frequency were matched between the two groups, the teaching intensity and interaction patterns differed substantially. The experimental group received real-time interactive feedback from the AI system and immediate manual corrections from the supervising teacher, whereas the control group received only traditional case discussions with delayed teacher summarization. This inherent imbalance in feedback immediacy and interaction frequency may have contributed to the observed differences in learning outcomes. We acknowledge that the superior performance of the experimental group may reflect not only the AI tool itself but also the cumulative effect of more intensive interactive engagement and immediate feedback. The experimental group has a natural advantage in interaction frequency and immediacy of feedback, and the observed differences in outcomes may be partly due to the variations in teaching intensity. Future studies should incorporate control conditions with matched interaction intensity to clarify the independent contribution of AI. Fifth, the presentation of imaging information is limited by the capabilities of the platform, and replacing real image interaction with textual descriptions may affect the authenticity of the teaching experience. In the future, multimodal AI technology could be integrated to enable combined image-text teaching. Finally, although the AI-simulated patient system provided real-time feedback and interactive training, potential risks inherent to large language model applications must be acknowledged. These risks include the possibility of model hallucinations (generation of clinically incorrect or unverified information), inaccurate or incomplete feedback on complex surgical decisions, and excessive learner dependence on AI-generated answers, which may undermine the development of independent clinical reasoning.

Conclusions

In this retrospective analysis of 157 neurosurgery residents, the AI-simulated patient teaching model based on a general-purpose large language model (DeepSeek) was associated with improved surgical decision-making ability. However, due to the retrospective, non-randomized design, causality cannot be established. The scores of the five dimensions in the experimental group were significantly higher than those in the traditional teaching group, including localization diagnosis, surgical approach selection, risk identification, emergency response, and surgical timing judgment. The excellent rate in each dimension was also higher, and student satisfaction was greater. With the continuous advancement of AI technology, the combination of intelligent simulation systems and human guidance is expected to become an important direction in surgical education.

Availability of Data and Materials

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Author Contributions

SR and JPZ designed the research study. SR, TG and SPR performed the research. JPZ, JFW and XMH analyzed the data. SR drafted the article. All authors contributed to the critical revision of the manuscript for important intellectual content. All authors read and approved the final

manuscript. All authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

Ethics Approval and Consent to Participate

This study obtained approval from the ethics committee of Taizhou Hospital of Zhejiang Province (approval number: K20260602) and was conducted in accordance with the Declaration of Helsinki. All participants signed a written informed consent.

Acknowledgment

Not applicable.

Funding

This research did not receive any funding.

Conflict of Interest

The authors declare no conflict of interest.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.62713/aic.4849.

References

[1] Rogers SO, Jr, Gawande AA, Kwaan M, Puopolo AL, Yoon C, Brennan TA, et al. Analysis of surgical errors in closed malpractice claims at 4 liability insurers. Surgery. 2006; 140: 25–33. https://doi.org/10.1016/j.surg.2006.01.008

[2] Wickramasinghe DP, Vincent J. The Use of Deliberate Practice in Simulation-Based Surgical Training for Open General and Subspeciality Surgery: A Systematic Review. Simulation in Healthcare : Journal of the Society for Simulation in Healthcare. 2026; 21: 298–305. https://doi.org/10.1097/SIH.0000000000000908

[3] Ma L, Yan R, Wang X, Gao X, Fan N, Liu L, et al. Enhancing Surgical Nursing Student Performance: Comparative Study of Simulation-Based Learning and Problem-Based Learning. Journal of Multidisciplinary Healthcare. 2024; 17: 991–1005. https://doi.org/10.2147/JMDH.S440333

[4] Roche AF, Kavanagh DO, Crimmins D, Healy V, Zilani G, Konge L, et al. A Delphi-based needs assessment to identify and prioritise procedural skills through consensus for simulation-based learning in neurosurgery. BMC Medical Education. 2025; 25: 328. https://doi.org/10.1186/s12909-025-06922-4

[5] Liu Y, Shi C, Wu L, Lin X, Chen X, Zhu Y, et al. Development and Validation of a Large Language Model-Based System for Medical History-Taking Training: Prospective Multicase Study on Evaluation Stability, Human-AI Consistency, and Transparency. JMIR Medical Education. 2025; 11: e73419. https://doi.org/10.2196/73419

[6] Zeng J, Qi W, Shen S, Liu X, Li S, Wang B, et al. Embracing the Future of Medical Education With Large Language Model-Based Virtual Patients: Scoping Review. Journal of Medical Internet Research. 2025; 27: e79091. https://doi.org/10.2196/79091

[7] McCarrick CA, McEntee PD, Boland PA, Donnelly S, O'Meara Y, Heneghan H, et al. A Randomized Controlled Trial of a Deep Language Learning Model-Based Simulation Tool for Undergraduate Medical Students in Surgery. Journal of Surgical Education. 2025; 82: 103629. https://doi.org/10.1016/j.jsurg.2025.103629

[8] Giglio B, Albeloushi A, Alhaj AK, Alhantoobi M, Saeedi R, Davidovic V, et al. Artificial Intelligence-Augmented Human Instruction and Surgical Simulation Performance: A Randomized Clinical Trial. JAMA Surgery. 2025; 160: 993–1003. https://doi.org/10.1001/jamasurg.2025.2564

[9] Ho CM, Guan S, Mok PKL, Lam CH, Ho WY, Mak CHK, et al. Development and Validation of a Large Language Model-Powered Chatbot for Neurosurgery: Mixed Methods Study on Enhancing Perioperative Patient Education. Journal of Medical Internet Research. 2025; 27: e74299. https://doi.org/10.2196/74299

[10] Redjem ID, Huaulmé A, Jannin P, Michinov E. Crisis management in the operating room: A systematic review of simulation training to develop non-technical skills. Nurse Education Today. 2025; 147: 106583. https://doi.org/10.1016/j.nedt.2025.106583

[11] Nakatani R, Patel K, Chowdhury T. Simulation in Anesthesia for Perioperative Neuroscience: Present and Future. Journal of Neurosurgical Anesthesiology. 2024; 36: 4–10. https://doi.org/10.1097/ANA.0000000000000939

[12] Society of Neurosurgery of Chinese Medical Association, Society of Cerebrovascular Surgery of Chinese Stroke Association, National Center for Neurological Disorders, National Clinical Research Center for Neurological Diseases. Chinese guideline for the clinical management of patients with ruptured intracranial aneurysms (2024). Zhonghua Yi Xue Za Zhi. 2024; 104: 1940–1971. https://doi.org/10.3760/cma.j.cn112137-20240222-00374

[13] Yao Z, Hu X, You C, He M. Effect and Feasibility of Endoscopic Surgery in Spontaneous Intracerebral Hemorrhage: A Systematic Review and Meta-Analysis. World Neurosurgery. 2018; 113: 348–356.e2. https://doi.org/10.1016/j.wneu.2018.02.022

[14] Lin V, Hutchinson PJ, Kolias A, Robba C, Wahlster S. Timing of neurosurgical interventions for intracranial hypertension: the intensivists' and neurosurgeons' view. Current Opinion in Critical Care. 2025; 31: 137–148. https://doi.org/10.1097/MCC.0000000000001243

[15] Hussain M, Hassan MM, Taj T, Shah V. A Cross-Sectional Study Assessing the Suitability of ChatGPT and DeepSeek AI for Generating Patient Education Guides on Imaging Modalities in Stroke. Cureus. 2025; 17: e92066. https://doi.org/10.7759/cureus.92066

© 2026 The Author(s).

