

JOURNAL OF TAZEEZ FOR PUBLIC HEALTH

AN OFFICIAL JOURNAL OF SAUDI HEALTH PROMOTION AND EDUCATION ASSOCIATION

Systematic review

AI-Enabled automated feedback and debriefing in simulation-based healthcare education: a systematic review of educational effectiveness

Dr. Saeed Abdullah Alzahrani

Pediatric Emergency Consultant, Pediatric Emergency Department, Prince Sultan Military Medical City, Riyadh, Saudi Arabia.

Email: dr.saeed.za@gmail.com

Abstract

Artificial intelligence is integrated into healthcare simulation to deliver individualized feedback and support reflective learning. Educational evaluations include procedural coaching, virtual patient encounters, and facilitator development. This review examined educational effectiveness while distinguishing observed performance from perceived competence and feedback feasibility. Original research published to May 2026 evaluating artificial intelligence enabled feedback or debriefing in healthcare simulation was reviewed. Comparative educational studies and complementary feedback validation or feasibility investigations were synthesized narratively. Outcomes included technical performance, clinical reasoning, communication, perceived competence, cognitive workload, and debriefing utility. Ten original reports comprised seven randomized trials and three prospective, pilot, or qualitative investigations. Surgical trials show improvements in selected performance measures, with benefits dependent on feedback timing, instructional design, and human participation. Personalized expert instruction informed by artificial intelligence exceeded artificial intelligence tutoring alone in practice and transfer tasks. Virtual patient feedback improved clinical reasoning in a small trial, while nursing simulations demonstrated gains in communication and perceived competence. Artificial intelligence enabled feedback offers benefits for simulation outcomes. Data for autonomous team debriefing is less developed than data for individualized coaching. Integration with educators, independent assessment, and longitudinal evaluation provides the strongest direction for educational implementation.

Keywords: Artificial intelligence; automated feedback; debriefing; simulation based education; healthcare education

Published: 1 Oct, 2026

<https://doi.org/10.65759/jwrsv661>

Copyright © 2025 The Author(s). Published by Lizzy B. This is an open-access article distributed under the terms of the Creative Commons Attribution (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited).

Introduction

Simulation based education provides structured opportunities to develop clinical knowledge, procedural skills, and professional performance within controlled learning environments (1). A large systematic review show improvements in multiple educational outcomes when technology enhanced simulation was compared with no intervention, establishing a foundation for its use in healthcare training (1). The educational value of simulation depends on instructional design rather than technological sophistication alone, with feedback, repetitive practice, curriculum integration, and explicit learning objectives identified as important features of effective programs (2). Deliberate practice extends this approach through repeated task performance, focused correction, and progressive improvement against defined standards, with comparative data supporting gains in clinical skill acquisition over traditional education (3).

Debriefing connects simulated experience with reflection by examining actions, decisions, and the reasoning underlying performance. It differs from isolated corrective feedback through its emphasis on shared interpretation, learner perspectives, and the application of lessons to future practice (4). A systematic review of technology enhanced simulation found favorable educational outcomes for simulation incorporating debriefing against no intervention, alongside variable findings across particular debriefing formats and design features (5).

Artificial intelligence introduces methods for analyzing performance data and translating identified errors into individualized instructional guidance. The Virtual Operative Assistant established an explainable framework linking machine learning assessment to educational feedback during simulated neurosurgical training

(6). Subsequent development of the Intelligent Continuous Expertise Monitoring System extended assessment to brief time intervals and incorporated coaching and risk detection using deep neural networks (7).

The emergence of automated instructional systems raises a specific educational question concerning the relationship between accurate assessment and effective learning (6,7). A tool that recognizes expertise or identifies errors still requires evaluation of its influence on subsequent performance, retention, and the transfer of skills to different tasks (3,7). Original articles analysis requires attention to the instructional intervention, comparator, measurement approach, and educational context rather than grouping all artificial intelligence applications as equivalent (2,5). This review examined original research on artificial intelligence enabled automated feedback and debriefing in simulation based healthcare education, prioritizing measured educational effectiveness and interpreting feedback validity, acceptability, and implementation findings as complementary data within an explicitly differentiated framework for informing educational practice and future evaluation.

Methods

This review was structured according to the Preferred Reporting Items for Systematic Reviews and Meta Analyses 2020 framework (Fig 1). The review question addressed healthcare learners and educators undertaking simulation, AI enabled feedback or debriefing, relevant educational comparators, and educational outcomes. Eligibility encompassed original quantitative, qualitative, and mixed methods research published in English from database inception through 31 May 2026.

PubMed, Scopus, Web of Science Core Collection, and the Cochrane Central Register of Controlled Trials were searched. Search concepts combined artificial intelligence, machine learning, deep learning, large language models, intelligent tutoring, automated feedback, debriefing, simulation, virtual patients, medical education, nursing education, and health professions education. Database specific syntax and controlled vocabulary were applied where available, and reference lists of relevant publications were examined for additional records. Duplicate records were removed before screening.

Two authors screened titles and abstracts against the eligibility criteria, and evaluated relevant full texts, recorded exclusion reasons, and resolved disagreements through discussion. Studies were eligible when artificial intelligence generated, selected, or informed feedback following learner performance, supported debriefing, or evaluated the educational validity or feasibility of such feedback.

Two authors independently extracted study design, setting, participant characteristics, sample size, simulation modality, artificial intelligence system, feedback timing, comparator, outcomes, and

numerical findings using a standardized form. Randomized effectiveness studies were distinguished from observational validation and feasibility investigations. Secondary reports from existing cohorts were linked to their primary publications to avoid duplicate representation. Observed technical performance, clinical reasoning, and communication were prioritized, while perceived competence, satisfaction, workload, feedback agreement, and implementation experiences were treated as secondary outcomes.

Two authors independently appraised randomized trials using the revised Cochrane risk of bias tool, with attention to randomization, deviations, missing data, outcome measurement, and selective reporting. Observational and qualitative investigations were assessed using design appropriate critical appraisal criteria, and disagreements were resolved by consensus. Findings were organized by procedural coaching, virtual patient feedback, and debriefing support. Narrative synthesis considered outcome direction, effect estimates, confidence intervals, comparator intensity, and assessment independence. Statistical pooling was not undertaken because interventions, measurement instruments, and educational contexts differed substantially.

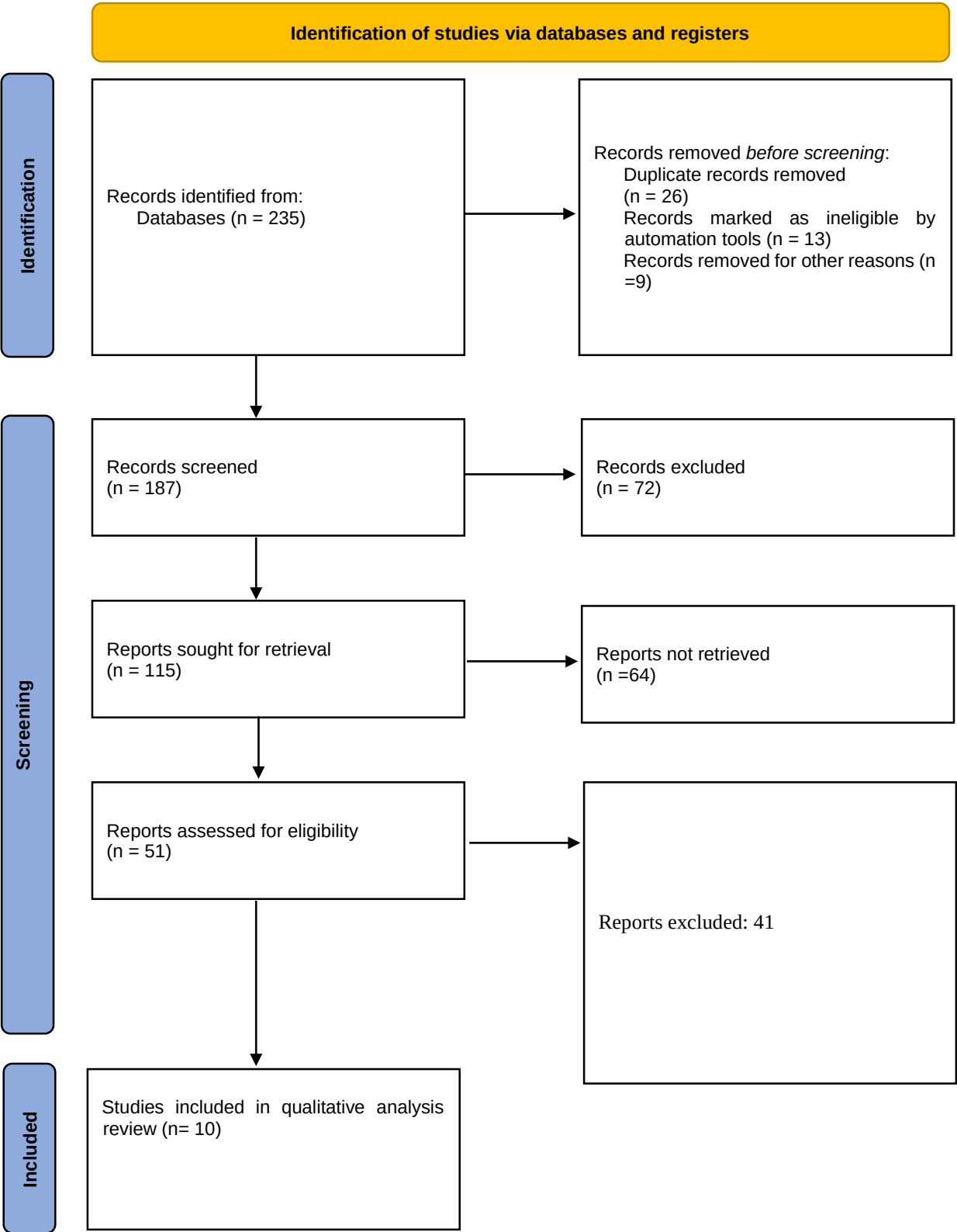


Figure 1: PRISMA flow chart

Results

Ten original reports comprised seven randomized trials and three investigations of feedback validity, feasibility, or user experience (8–17). Four trials examined simulated neurosurgical training, while three evaluated medical history taking or nursing simulation outcomes (8–12,14,15). The remaining studies assessed automated history feedback, facilitator feedback, and teamwork reports, providing complementary process data rather than comparative learning effects (13,16,17) (Table 1).

In the seventy student trial by Fazlollahi et al., the Virtual Operative Assistant improved practice expertise scores against remote expert instruction by 0.66 points, with a 95% confidence interval of 0.55–0.77. Scores during a more realistic simulated resection also favored artificial intelligence tutoring, supporting transfer within the simulation environment (8). Yilmaz et al. analyzed ninety-seven learners and found better final practice composite scores with real time artificial intelligence feedback than with in person instruction, alongside similar overall expert rated technical scores between those groups. The realistic task composite outcome did not differ significantly, illustrating variation across performance measures (9).

In Giglio et al., eighty-eight students were randomized and eighty-seven were analyzed, with personalized expert instruction informed by artificial intelligence outperforming intelligent tutoring alone. The mean difference was 0.26 points during practice trial five and 0.20 during the realistic transfer task (10). A separate twenty-five student crossover trial show a sequence effect: expert instruction followed by artificial intelligence produced greater improvement than the reverse sequence, with a reported mean difference of 0.67 points (11). These findings supported integrated instruction without establishing universal superiority of autonomous tutoring (10,11).

Brügge et al. analyzed twenty one medical students practicing histories with simulated patients and found higher final clinical reasoning scores with structured artificial intelligence feedback than without feedback. Final scores were 3.60 versus 3.02, with a group by time effect reaching statistical significance and partial eta squared of 0.198 (12). Holderried et al. evaluated one hundred six simulated conversations and reported strong agreement between automated and human feedback, with Cohen's kappa of 0.832. This agreement established feedback classification validity within the tested setting, without demonstrating improvement in subsequent learner performance (13).

Fung et al. enrolled forty-four nursing students and reported greater initial improvement in perceived competence after generative artificial intelligence patient simulation than after immersive video simulation. Both interventions incorporated facilitator debriefing, and the artificial intelligence condition additionally provided individualized automated feedback (14). Dai et al. studied sixty-five nursing students and reported gains of 3.38 communication points and 3.91 percentage points in decision accuracy, together with shorter task completion time (15).

Vaughn et al. evaluated a retrieval augmented feedback chatbot with six facilitators, establishing feasibility through a posttest usability design (16). Tscholl et al. examined twenty six teamwork reports and twenty seven learner interviews, identifying useful observations alongside categorization and speaker attribution errors. Learners did not receive the generated reports, precluding interpretation as a learner debriefing effectiveness trial (17). Across the data set, outcome heterogeneity and limited follow up restricted conclusions about sustained competence and workplace transfer, without establishing effects on patient outcomes (8–17) (Table 2).

Table 1. Characteristics of the included studies

Study	Design and sample	Setting	Feedback and comparator
Fazlollahi et al., 2022	Randomized trial; 70 medical students	Canada; neurosurgical VR	VOA feedback versus remote expert instruction or practice (8)
Yilmaz et al., 2024	Randomized trial; 97 analyzed	Canada; neurosurgical VR	Real time ICEMS versus expert instruction or posttask feedback (9)
Giglio et al., 2025	Randomized trial; 88 randomized, 87 analyzed	Canada; neurosurgical VR	AI informed personalized expert instruction versus AI or scripted instruction (10)
Yilmaz et al., 2025	Crossover trial; 25 students	Canada; neurosurgical VR	Expert then AI instruction versus reverse sequence (11)
Brügge et al., 2024	Randomized trial; 21 analyzed	Germany; virtual histories	ChatGPT feedback versus identical patient simulation without feedback (12)
Holderried et al., 2024	Prospective validation; 106 conversations	Germany; virtual histories	GPT-4 feedback compared with human ratings (13)
Fung et al., 2025	Crossover trial; 44 nursing students	International undergraduate program	GenAI patient simulation with additional feedback versus 360° VR; both received facilitator debriefing (14)
Dai et al., 2026	Crossover trial; 65 nursing students	China; VR OSCEs	ChatGPT immediate feedback versus VR alone (15)
Vaughn et al., 2025	Posttest pilot; six facilitators	United States; facilitator development	SIM-U retrieval augmented debriefing feedback; no effectiveness comparator (16)
Tscholl et al., 2026	Qualitative study; 41 participants, four reviewers	Switzerland; team simulation	AI teamwork reports assessed by experts; learners did not receive reports (17)

Table 2. Educational findings and interpretation

Study	Principal finding	Educational interpretation
Fazlollahi et al.	Practice expertise difference versus expert: 0.66; 95% CI 0.55–0.77	Benefit versus remote instruction on the tested simulator (8)
Yilmaz et al., 2024	Final practice difference versus expert: 0.332; 95% CI 0.173–0.491	Overall OSATS similar between AI and expert groups (9)
Giglio et al.	Personalized expert versus AI: practice difference 0.26; transfer difference 0.20	Human interpretation of AI data improved learning (10)
Yilmaz et al., 2025	Expert then AI sequence advantage: 0.67; 95% CI 0.43–0.91	Instructional order influenced effectiveness (11)
Brügge et al.	Final CRI-HTI: feedback 3.60 versus control 3.02; $p=0.049$	Improvement in observed reasoning during simulated histories (12)
Holderried et al.	Automated versus human feedback agreement: $\kappa=0.832$	Feedback validity demonstrated; learning gain not tested (13)
Fung et al.	Initial adjusted CCQ difference: 16.59; 95% CI 2.77–30.41	Self-reported competence; combined modality and feedback intervention (14)
Dai et al.	Communication +3.38 points; decision accuracy +3.91 percentage points; completion time –0.78 minutes	Immediate simulation performance improved (15)
Vaughn et al.	Feasible feedback chatbot evaluated through facilitator usability survey	Supports implementation feasibility; sustained effectiveness untested (16)
Tscholl et al.	Useful observations alongside speaker attribution and categorization errors	Supports supervised development; learner effectiveness untested (17)

Discussion

Our systematic review findings support educational benefits from artificial intelligence enabled feedback across selected procedural and communication outcomes, with substantial dependence on instructional design and comparator choice. The strongest comparative findings concerned repeated practice with individualized correction rather than autonomous management of complete team debriefings (8–12,15). Interpretation rests on the specific educational task, the feedback delivered, and the independence of outcome assessment rather than on artificial intelligence as a uniform intervention (9,10,13).

A central finding was the continuing contribution of human educators within technically sophisticated learning systems (10,11). Personalized instruction informed by algorithmic error data improved practice and transfer performance, while crossover results favored expert teaching before subsequent artificial intelligence coaching (10,11). These observations align with a secondary analysis reporting that artificial intelligence guided curriculum design improved safety measures while adversely affecting selected efficiency measures, demonstrating the importance of balancing educational targets (18). Together, the findings support educator involvement in defining competencies, interpreting errors, and addressing performance dimensions outside the automated system's selected metrics (10,18).

Debriefing represents a broader educational activity than delivery of corrective statements or numerical performance scores (4,19). The PEARLS framework integrates learner self assessment, focused facilitation, and directive feedback or teaching, providing a useful conceptual basis for positioning automated outputs within a facilitated discussion (19). Debriefing with good judgment similarly emphasizes examining the reasoning behind actions

through advocacy and inquiry, a function that performance classification alone does not establish (20). TeamGAINS further addresses team interaction dynamics through structured reflection, reinforcing the distinction between generating a teamwork report and facilitating collective learning (21).

The communication studies provide complementary data for automated feedback beyond surgical psychomotor training (12,13,15). Improved history taking scores and nursing communication outcomes indicate that structured feedback contributes to learning within virtual encounters (12,15). Agreement between automated and human feedback strengthens confidence in the tested classification process, although agreement and educational benefit remain different outcomes requiring separate evaluation (13). Perceived competence gains in the nursing crossover study also reflected a combined intervention containing different simulation modalities and additional artificial intelligence feedback, preventing attribution of the entire effect to feedback alone (14). These distinctions are essential when translating findings into curricula (14).

Implementation data identified useful features with important limitations in contextual interpretation and usability (16,17). The wider generative artificial intelligence simulation literature describes diverse applications and outcomes, supporting careful differentiation between technical feasibility, learner satisfaction, and demonstrated effectiveness (22). Errors in speaker attribution and categorization carry educational significance because inaccurate observations alter the basis of subsequent discussion and feedback (17). Concentration of surgical trials within one research program, small cohorts, variable comparators, and short observation periods restrict broader generalization (8–12).

Future evaluations require transparent reporting of system versions, prompts, instructional content, feedback timing, and the contribution of human facilitation. Objective assessments across unfamiliar scenarios and longer follow up are necessary to establish retention and transfer beyond the immediate training task. Trials that compare equivalent practice exposure and separate automated feedback from changes in simulation modality offer clearer estimates of its educational contribution.

Conclusion

Artificial intelligence enabled automated feedback improved educational outcomes in healthcare simulation, particularly technical performance, clinical reasoning, and communication. Benefits varied across systems, comparators, and instructional sequences, with several findings favoring integration of artificial intelligence and human teaching. Feedback validation and feasibility studies supported further development without establishing sustained learning effects. Educational implementation requires clear learning objectives, independent performance assessment, and accountable educator involvement.

List of abbreviations

AI	Artificial intelligence
CCQ	Clinical Competence Questionnaire
CI	Confidence interval
CRI-HTI	Clinical Reasoning Indicator–History Taking Inventory
GenAI	Generative artificial intelligence
GPT	Generative Pretrained Transformer
ICEMS	Intelligent Continuous Expertise Monitoring System
OSATS	Objective Structured Assessment of Technical Skills

List of abbreviations

OSCE	Objective Structured Clinical Examination
PEARLS	Promoting Excellence and Reflective Learning in Simulation
VOA	Virtual Operative Assistant
VR	Virtual reality

References

1. Cook DA, Hatala R, Brydges R, Zendejas B, Szostek JH, Wang AT, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. **JAMA**. 2011;306(9):978–988.
2. Issenberg SB, McGaghie WC, Petrusa ER, Gordon DL, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. **Med Teach**. 2005;27(1):10–28.
3. McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. **Acad Med**. 2011;86(6):706–711.
4. Fanning RM, Gaba DM. The role of debriefing in simulation-based learning. **Simul Healthc**. 2007;2(2):115–125.
5. Cheng A, Eppich W, Grant V, Sherbino J, Zendejas B, Cook DA. Debriefing for technology-enhanced simulation: a systematic review and meta-analysis. **Med Educ**. 2014;48(7):657–666.
6. Mirchi N, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, Del Maestro RF. The Virtual Operative Assistant: an explainable artificial intelligence tool for simulation-based training in surgery and medicine. **PLoS One**. 2020;15(2):e0229596.
7. Yilmaz R, Winkler-Schwartz A, Mirchi N, Reich A, Christie S, Tran DH, et al. Continuous monitoring of surgical bimanual

- expertise using deep neural networks in virtual reality simulation. **NPJ Digit Med.** 2022;5:54.
8. Fazlollahi AM, Bakhaidar M, Alsayegh A, Yilmaz R, Winkler-Schwartz A, Mirchi N, et al. Effect of artificial intelligence tutoring vs expert instruction on learning simulated surgical skills among medical students: a randomized clinical trial. **JAMA Netw Open.** 2022;5(2):e2149008.
 9. Yilmaz R, Bakhaidar M, Alsayegh A, Abou Hamdan N, Fazlollahi AM, Tee T, et al. Real-Time multifaceted artificial intelligence vs In-Person instruction in teaching surgical technical skills: a randomized controlled trial. **Sci Rep.** 2024;14:15130.
 10. 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 Surg.** 2025;160(9):984–991.
 11. Yilmaz R, Alsayegh A, Bakhaidar M, Fazlollahi AM, Abou Hamdan N, Tee T, et al. Combining real-time AI and in-person expert instruction in simulated surgical skills training—randomized crossover trial. **NPJ Artif Intell.** 2025;1:36.
 12. Brügge E, Ricchizzi S, Arenbeck M, Keller MN, Schur L, Stummer W, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. **BMC Med Educ.** 2024;24:1391.
 13. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Festl-Wietek T, Holderried M, Eickhoff C, et al. A language model–powered simulated patient with automated feedback for history taking: prospective study. **JMIR Med Educ.** 2024;10:e59213.
 14. Fung TCJ, Chan SL, Lam CFM, Lam CY, Cheng CCW, Lai MH, et al. Effects of generative artificial intelligence (GenAI) patient simulation on perceived clinical competency among global nursing undergraduates: a cross-over randomised controlled trial. **BMC Nurs.** 2025;24:934.
 15. Dai X, Gong Y, Ma LY. Effects of ChatGPT-generated immediate feedback integrated into VR-based OSCEs on nursing students' performance: a randomized crossover study. **BMC Nurs.** 2026;25:396.
 16. Vaughn J, Ford SH, Sheckells A, Fasnacht H, Crawford D, Dogan G. Enhancing simulation facilitator debriefing using a generative artificial intelligence feedback interface and retrieval-augmented generation: a pilot study. **Clin Simul Nurs.** 2025;109:101843.
 17. Tscholl DW, Ebensperger M, et al. Generative AI in simulation debriefings: an exploratory study using the Team-FIRST framework and qualitative feedback from simulation experts and learners. **Adv Simul (Lond).** 2026;11:14.
 18. Fazlollahi AM, Yilmaz R, Winkler-Schwartz A, Mirchi N, Ledwos N, Bakhaidar M, et al. AI in surgical curriculum design and unintended outcomes for technical competencies in simulation training. **JAMA Netw Open.** 2023;6(9):e2334658.
 19. Eppich W, Cheng A. Promoting Excellence and Reflective Learning in Simulation (PEARLS): development and rationale for a blended approach to health care simulation debriefing. **Simul Healthc.** 2015;10(2):106–115.
 20. Rudolph JW, Simon R, Dufresne RL, Raemer DB. There's no such thing as "nonjudgmental" debriefing: a theory and method for debriefing with good judgment. **Simul Healthc.** 2006;1(1):49–55.
 21. Kolbe M, Weiss M, Grote G, Knauth A, Dambach M, Spahn DR, et al. TeamGAINS: a tool for structured debriefings for simulation-based team trainings. **BMJ Qual Saf.** 2013;22(7):541–553.

Neo NWS, Gunawan J, Levett-Jones T, Khoo ET, Chua WL, Liaw SY. Generative artificial intelligence in healthcare simulation-based education: a scoping review. **Clin Simul Nurs.** 2025;108:101819.