PRiMER
peer-reviewed reports in medical education research

research brief

Comparative Outcomes of ChatGPT Versus Faculty-Facilitated Simulation Design by Medical Students

Chad N. Taniguchi, MD | Keely Myers, MD | Erin Jyo, MD | Kyle Ishikawa, MS | Benjamin W. Berg, MD | Jannet Lee-Jayaram, MD

PRiMER. 2026;10:32.

Published: 09/10/2026 | DOI:10.22454/PRiMER.2026.243533

Abstract

Introduction: Limited research describes generative artificial intelligence (AI) and large language models, such as ChatGPT, to support the design of health care simulation-based manikin scenarios (SBMS). This pilot study describes the design process and quality assessment of SBMS designed by medical students with ChatGPT generative AI support (AIG scenarios) compared to those designed by medical student interest groups (MSIG) with faculty guidance (MSIG scenarios).

Methods: Based on this medical school simulation center's scenario design template, we compared the time to completion of the scenario between five MSIG SBMS and five AIG SBMS on the same clinical topics. Three blinded simulation experts rated all randomized scenarios using the Simulation Scenario Evaluation Tool (SSET). We performed paired *t* tests to evaluate differences between the MSIG and AIG scenarios.

Results: MSIG SBMS design required months of faculty-student collaboration, whereas AIG SBMS were completed in a mean of 37 minutes. AIG scenarios had significantly higher Scenario Materials and Resources SSET scores. We found no differences in other SSET elements or overall scores. Rater agreement across most categories was low. The Debriefing Plan element received the lowest scores in both groups.

Conclusions: This study supports the potential of simulation novices using ChatGPT to streamline health care SBMS design. AIG scenarios required less time to create and were comparable in quality to MSIG scenarios. By accelerating the initial scenario design process, AI offers an efficient tool for developing educational simulations, expanding opportunities for students to engage in scenario design as an educational methodology.

Introduction

Interest is growing in applying generative AI models, such as ChatGPT, to simulation-based health care education, which is a mainstay in medical education for developing knowledge and skills in a safe learning environment.[1-4] Simulation-based manikin scenarios (SBMS) replicate various patient conditions, allowing learners to apply clinical knowledge and practice problem-solving, decision-making, and teamwork in a variety of specialties, across acute and primary care settings. Cocreating educational content with faculty guidance enhances learner engagement, promotes ownership of learning, and reinforces knowledge.[5] It also allows faculty to improve teaching practices, gain insight into learner needs, and foster an inclusive learning environment.[5]

Existing studies examining ChatGPT's role in SBMS design have focused on the use of AI by simulation experts or nonsubject matter experts (SME) demonstrating efficiency gains while identifying limitations in accuracy and content quality, but none have evaluated its application by novice simulationists.[4,6] Given their inexperience in scenario design, novice simulationists stand to benefit most from AI-assisted tools. However, unlike experts, they may lack the knowledge to identify inaccurate or suboptimal AI-generated content, making quality evaluation of AI-assisted scenario design particularly important in the population. This pilot study examined the ChatGPT-assisted design process and compared novice-designed ChatGPT-assisted generative AI (AIG) SBMS with faculty-facilitated medical student interest groups (MSIG) SBMS.

Methods

The University of Hawaii Office of Research Compliance determined this study not to be human subject research (Protocol 2026–00063). Five MSIG-designed scenarios on open femur fracture, preeclampsia, tricyclic antidepressant intoxication, alcohol withdrawal, and epidural hematoma were selected from the medical school's simulation center archives. MSIG are specialty-focused extracurricular student organizations that can design and run supplemental simulation-based student events. MSIG scenarios were designed by medical students with no simulation design experience, under unstructured guidance from specialty and simulation faculty, through iterative in-person meetings and email communications over the course of months. All MSIG scenarios followed the simulation center's standardized, 18-component design template (Appendix 1) and were used in extracurricular MSIG simulations.

Three second-year medical student scenario designers (MSSD) created AIG scenarios on the same MSIG SBMS topics. MSSD had participated as learners in standard curricular SBMS but had no prior experience in scenario design, were not involved in the MSIG scenario design, nor had they participated in any of the MSIG SBMS. MSSD accessed MSIG scenario synopses and learning objectives only, without viewing the full MSIG scenario design templates. AIG-assisted scenario design began with an initial prompt in ChatGPT v3.5 (OpenAI) based on the MSIG scenario synopsis (Appendix 2).[7] MSSD iteratively refined prompts until sufficient information was generated to complete the scenario design template. No files were uploaded with the prompts, no custom generative pretrained transformers (GPTs) were used, and no faculty assistance was provided during the AIG design process.

Outcome Measures

Three external physician faculty with simulation expertise rated the 10 scenarios in a unique, randomized sequence. Raters were blinded to the study aims and scenario group. Scenarios were provided in the identical scenario design template for both groups, though scenario length was not formally assessed. Raters used the Simulation Scenario Evaluation Tool (SSET) to assess scenario quality.[8] The SSET included six elements with a total of 20 items: Learning Objectives, Clinical Context/Scenario Overview, Critical Actions, Patient States, Scenario Materials and Resources, and Debriefing Plan. Each element had two to seven items, scored on a scale of 1 (low) to 5 (high). Raters were provided three practice scenarios, separate from the study scenarios, to familiarize themselves with the SSET. No additional rater training was provided.

Data Analysis

We calculated the mean scenario generation time and number of prompts for the AIG scenarios. We used paired *t* tests and Wilcoxon signed-rank tests to compare SSET scores between groups. We calculated the intraclass correlation coefficient (ICC) for interrater reliability with the intent that the rater's score would be *averaged* per scenario. We specified a *two-way* random effects model because a fixed number of raters score a fixed set of scenarios. Statistical significance was set at $P<0.05$. All analyses were performed in R version 4.3.0 (R Foundation).

Results

MSIG scenario design required multiple faculty-student meetings over the span of months, along with unmeasured student time. AIG design time was 36.6±18.27 (mean ± SD) minutes, using a mean of 5.8±1.3 prompts per scenario. The individual SSET mean element score in Scenario Materials and Resources was significantly higher in AIG scenarios (3.83±0.20) compared to MSIG scenarios (2.67±0.70; $P=0.017$). We found no statistically significant differences between AIG and MSIG SSET overall or other element scores; however, AIG scenarios demonstrated higher mean scores overall and across all elements (Table 1).

Table 1. AIG and MSIG Scenario SSET Scores

	AIG, *N* = 5 Mean (SD)	MSIG, *N* = 5 Mean (SD)	Difference*	*P* value*	Agreement	Consistency
Overall score	3.49 (0.33)	3.10 (0.48)	0.39	.245	0.530	0.507
Element 1: Learning objectives	3.36 (0.49)	3.30 (0.47)	0.07	.806	0.340	0.369
Element 2: Clinical content/Scenario Overview	3.43 (0.22)	3.23 (0.45)	–	.584	−1.748	−1.402
Element 3: Critical actions	3.64 (0.41)	3.31 (0.40)	0.33	.107	0.223	0.265
Element 4: Patient states	3.85 (0.45)	3.20 (0.47)	0.65	.146	0.227	0.295
Element 5: Scenario materials and resources	3.83 (0.20)	2.67 (0.70)	1.17	.017	0.494	0.475
Element 6: Debriefing plan	2.70 (0.61)	2.37 (1.27)	–	.625	0.760	0.759

*Paired *t* test; Wilcoxon signed rank test with continuity correction; Wilcoxon signed rank exact test.
Abbreviations: AIG, generative artificial intelligence group; MSIG, medical student interest group; SD, standard deviation; SSET, Simulation Scenario Evaluation Tool

The SSET Debriefing Plan element mean scores were lowest for both groups: 2.70±0.61 for AIG and 2.37±1.27 for MSIG. Clinical Content/Scenario Overview demonstrated negative ICC values for agreement (−1.748) and consistency (−1.402), demonstrating a strong disagreement. Within each scenario, the variability between scores was greater than the variability within each rater, leading to large negative ICC values (<−

1.0). Debriefing Plan had moderate interrater reliability with an ICC of 0.760 for agreement and 0.759 for consistency. All other elements had low interrater reliability (ICC<0.50).

Discussion

This pilot study provides support for ChatGPT-assisted novice scenario design. ChatGPT-assisted scenarios surpassed the traditional design process in time-based efficiency. AIG scenarios received higher mean overall scores and in every SSET element; however, only the Scenario Materials and Resources element reached statistical significance. This consistent directionality may represent a signal that the study was underpowered to detect. For example, although the MSSD incorporated MSIG learning objectives into the prompt, the AI also was instructed to generate learning objectives, potentially resulting in clearer or more structured wording, contributing to the differences in Learning Objectives element scores between groups. The findings suggest that novices assisted by AI can generate SBMS with scenario quality comparable to, and possibly higher than, SBMS with student-faculty collaboration. Generative AI may provide a useful starting point for simulation scenario development, especially for novices. However, human SME oversight remains crucial to confirm clinical accuracy and alignment with learning objectives.[9]

Both AIG and MSIG scenarios received their lowest SSET scores in the Debriefing Plan element. AIG debriefing plans consisted of open-ended questions, which may be insufficient to capture the complexity of effective debriefing. These suboptimal ratings suggest that ChatGPT prompts may require enhancement to produce more comprehensive debriefing plans, potentially by incorporating structured frameworks, such as PEARLS or GAS, which are designed to guide debriefers in applying best practices following simulation.[10,11]

Student-led SBMS design may serve as an educational activity. Student-led scenario design is underexplored, although a few studies have suggested enhanced learning outcomes for student designers.[12,13] Engaging students in both specialty and primary care scenario design may reinforce traditional learning, fostering a deeper understanding of clinical concepts. Rather than replacing faculty collaboration, AI-assisted design may serve as a starting point from which novice designers or students can engage SME faculty at a more advanced stage, potentially enriching mentorship experiences and clinical reasoning development.

Limitations included a small sample size and a limited number of raters, which reduced statistical power, and the study was not designed or powered to test for equivalence or noninferiority. Findings should be interpreted as preliminary and require replication in larger samples before definitive conclusions can be drawn. Lack of formal rater training and calibration may have contributed to low interrater reliability across SSET elements, including negative ICC values reflecting disagreement among raters. The SSET had limited prior usage and no threshold for defining scenario quality.[14] MSIG scenario design time was not formally measured; the differing methods used to capture design time represent a further limitation. Learning outcomes for MSSD were not evaluated, limiting conclusions about the utility of AI-assisted scenario creation as a learning activity. Future studies should explore the educational impact of AI-assisted scenario design, accounting for AI output variability, and the broader ethical implications of AI use in medical education.

Conclusions

This pilot study supports the potential of simulation novices using AI to streamline the initial SBMS design process. AIG scenarios required less time to create and were comparable in quality to MSIG scenarios, with consistently higher mean scores that did not reach statistical significance. AI offers an efficient tool for making SBMS creation more accessible to students and novice educators, while expanding opportunities for simulation-based learning. The efficiency afforded during initial development may allow novice designers to focus on other critical elements for optimal learning and collaboration with expert simulationists and SMEs.

Acknowledgments

We thank the simulation experts who participated in this project: Dr. Ismail Saiboon, Dr. Daniel Salcedo, and Dr. Len Tanaka.

Support

The project described was supported by the JABSOM Office of the Dean. Kyle M. Ishikawa was partially supported by the U54MD007601 (Ola HAWAII) grant from the National Institute of Health (NIH). The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

Presentations

Hawaii Professions Education Conference, February 10, 2024, Honolulu, HI; International Meeting on Simulation in Healthcare, January 13, 2025, Orlando, FL

Corresponding Author

Chad N. Taniguchi

John A. Burns School of Medicine, Honolulu, HI

chad73@hawaii.edu

Author Affiliations

Chad N. Taniguchi, MD - John A. Burns School of Medicine, Honolulu, HI

Keely Myers, MD - John A. Burns School of Medicine, Honolulu, HI

Erin Jyo, MD - John A. Burns School of Medicine, Honolulu, HI

Kyle Ishikawa, MS - Department of Quantitative Health Sciences, John A. Burns School of Medicine, Honolulu, HI

Benjamin W. Berg, MD - SimTiki Simulation Center, John A. Burns School of Medicine, Honolulu, HI

Jannet Lee-Jayaram, MD - SimTiki Simulation Center, John A. Burns School of Medicine, Honolulu, HI

References

1. Alinier G, Oriot D. Simulation-based education: deceiving learners with good intent. *Adv Simul (Lond)*. 2022;7(1):8. doi:10.1186/s41077-022-00206-3
2. Motola I, Devine LA, Chung HS, Sullivan JE, Issenberg SB. Simulation in healthcare education: a best evidence practical guide. AMEE Guide No. 82. *Med Teach*. 2013;35(10):e1511–30. doi:10.3109/0142159X.2013.818632
3. Harder N. Advancing healthcare simulation through artificial intelligence and machine learning: exploring innovations. *Clinical Simulation in Nursing*. 2023;83:101456. doi:10.1016/j.ecns.2023.101456
4. Rodgers DL, Needler M, Robinson A, et al. Artificial intelligence and the simulationists. *Simul Healthc*. 2023;18(6):395–399. doi:10.1097/SIH.0000000000000747
5. Könings KD, Mordang S, Smeenk F, Stassen L, Ramani S. Learner involvement in the co-creation of teaching and learning: AMEE Guide No. 138. *Med Teach*. 2021;43(8):924–936. doi:10.1080/0142159X.2020.1838464

6. Violato E, Corbett C, Rose B, Rauschning B, Witschen B. The effectiveness and efficiency of using ChatGPT for writing health care simulations. *ijohs*. 2023. doi:10.54531/wjgb5594
7. OpenAI. Platform. Accessed September 22, 2026. https://platform.openai.com
8. Hernandez J, Frallicciardi A, Nadir N-A, Gothard MD, Ahmed RA. Development of a simulation scenario evaluation tool (SSET): modified Delphi study. *BMJ Simul Technol Enhanc Learn*. 2020;6(6):344–350. doi:10.1136/bmjstel-2019-000521
9. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930–1940. doi:10.1038/s41591-023-02448-8
10. 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. doi:10.1097/SIH.0000000000000072
11. Phrampus PE, O'Donnell JM. Debriefing using a structured and supported approach. In: 10.1007/978-1-4614-5993-4_6Levine AI, DeMaria S, Schwartz AD, Sim AJ, eds. *The Comprehensive Textbook of Healthcare Simulation*. Springer; 2013:73–84 .
12. Steinemann SK, Berg BW, Turban JW, Hara KM, Alfrey LR. Student-written simulation scenarios: a novel cognitive assessment method in a trauma curriculum. *Hawaii Med J*. 2011;70(8):172–175. https://pmc.ncbi.nlm.nih.gov/articles/PMC3158380
13. Gwin T, Villanueva C, Wong J. Student developed and led simulation scenarios. *Nurs Educ Perspect*. 2017;38(1):49–50. doi:10.1097/01.NEP.0000000000000080
14. Rodgers DL, Hernandez J, Ahmed RA. Response to Bhutiani, Hester, and Lonsdale. *Simul Healthc*. 2024;19(4):270. doi:10.1097/SIH.0000000000000817