Clinical Simulation in Nursing 110 (2026) 101883

Contents lists available at ScienceDirect

Clinical Simulation in Nursing

journal homepage: www.elsevier.com/locate/ecns

Featured Article

Using AI to create simulation scenarios for a screening brief intervention and referral to treatment virtual reality simulation

Elizabeth Wells-Beede, PhD, RN, C-EFM, CHSE-A, CNE, ACUE, FSSH, FAAN[a,b,*], Lauren Thai, MEd, CHSOS[b], Jinsil Hwaryoung Seo, PhD[c], Mihir Sunil Godbole[c], Jay Hareshbhai Patel, BDS, MPH[a], Cindy Weston, DNP, APRN, FNP-BC, CHSE, FNAP, FAANP, FAAN[a], Nicole Kroll, PhD, APRN, ANP-C, FNP-BC, PMHNP-BC[d]

[a] University of North Texas Health Sciences Center College of Nursing, Ft. Worth, TX, USA
[b] Texas A&M University College of Nursing, Round Rock, TX, USA
[c] Texas A&M University College of Performing, Visual and Fine Arts, College Station, TX, USA
[d] Texas A&M College of Nursing, College Station, TX, USA

ARTICLE INFO

Keywords:
Artifiical intelligence
Referral
Screening
Virtual reality

ABSTRACT

A virtual reality (VR) platform for training nursing students in screening brief intervention referral to treatment (SBIRT) was limited by a single scenario. To expand this, we used an artificial intelligence (AI) chatbot (OpenAI's ChatGPT) to generate three new scenarios. The primary focus of this study was the content validation of this AI-generated content through a modified Delphi method. A panel of five (n = 5) certified subject matter experts (SMEs) evaluated the scenarios using the Simulation Scenario Evaluation Tool (SSET). The expert review revealed a critical divergence in consensus: SMEs reached "substantial agreement" ($\kappa = 0.80$) on the procedural "Critical Actions", but only "fair agreement" ($\kappa = 0.31$-0.36) on key educational components, such as the "Debriefing Plan". This study validates a "human-in-the-loop" model, demonstrating that while AI is a powerful tool for developing the core content of simulations, meticulous SME review and refinement remain essential for creating intellectually sound and effective educational experiences. In conclusion, these scenarios have achieved high overall ratings and are currently being integrated for pilot use.

Background

Screening, brief intervention, and referral to treatment (SBIRT) programs were established to provide early intervention for individuals identified with high-risk substance use (Substance Abuse and Mental Health Services Agency, 2011). This easy-to-implement public health model has been implemented to offer universal screening that can be incorporated in health care settings, including academia. Students in health-related programs can be trained on SBIRT for incorporation into practice. In 2023, it was reported that over 21.7% of people binge drink, over 5.8% are heavily consuming alcohol, and 24.9% used an illicit drug (2024). So, proactively providing educational screening opportunities can help identify patients who may need support due to high-risk substance use, defined by 2024 as consumption above established low-risk limits.

Yeh et al. (2024) demonstrated that virtual reality (VR) simulation significantly increased nursing students' confidence in conducting SBIRT. In their study, both undergraduate and graduate students reported higher self-efficacy scores after practicing SBIRT in a VR environment, and participants consistently identified VR as a promising educational tool for skill development. However, the VR platform evaluated in that research was limited to a single scenario. Specifically, the simulation included only a single case scenario involving a patient with alcohol misuse, which was the sole case available for use at the time of the study.

The initial VR scenario provided a useful SBIRT practice experience with a cooperative patient. However, after multiple uses, the team recognized the need to expand beyond this single case and the value of expanding SBIRT cases to cover other patient populations and settings such as psychiatric mental health, physician assistants and/or licensed mental health care providers. Through the support of Health Resources and Services Administration (HRSA) funds, the SBIRT VR team was able to design three new VR simulation scenarios. These new cases included scenarios addressing opi-

* Corresponding author.
E-mail address: elizabeth.beede@unthsc.edu (E. Wells-Beede).

https://doi.org/10.1016/j.ecns.2025.101883

oid use and exploring the possibility of medication-assisted treatment, among other variations, thereby broadening the platform's educational scope. We conducted a modified Delphi method to achieve consensus on content validity; a process modeled on the prior work of Hernandez et al. (2019). The VR simulation scenarios were assessed using the Simulation Scenario Evaluation Tool (SSET) to achieve consensus of content validity.

Method

Scenario development with ChatGPT

The research team utilized OpenAI's ChatGPT conversational artificial intelligence (AI) platform to generate three distinct VR simulation-based learning scenarios by entering specific parameter prompts (OpenAI, 2024). The original SBIRT VR scenario, created in 2020 and discussed in Yeh et al. (2024), was created manually. In contrast, this study aimed to simplify new scenario development and test AI's capabilities by using a standardized prompt template to guide ChatGPT in scenario generation.

Scenario creation began by asking ChatGPT to produce a concise scenario goal (i.e., the primary learning objective), followed by a request for a full simulation-based learning experience based on that goal. Each prompt specified the required scenario elements including learning objectives, patient demographics/background, a dialogue using SBIRT and screening questions, assessment cues, and debriefing points, in alignment with the Healthcare Simulation Standards of Best Practice: Design™ (INACSL Standards Committee et al., 2021). The prompts were structured to be specific and comprehensive, as shown in the following examples:

- One prompt given was: "Generate a detailed SBIRT simulation scenario for a nursing student. The patient is 20-year-old college student misusing prescription opioids. Include the scenario's goal, relevant patient history and demographics and realistic nurse patient dialogue incorporating SBIRT and CAGE-AID questions, and key debriefing points."
- In another case, we prompted: "Provide an SBIRT simulation scenario where the patient has alcohol use disorder for 6 years and is conservative and stubborn to change. Include clear learning objectives, the patient's context and vital information, a realistic conversation utilizing motivational interviewing, and important points for debriefing."

Then the AI-generated content for each scenario was transferred verbatim into a standard simulation scenario template document for further refinement by the team. Notably, we ensured that the CAGE-AID screening tool (Cut down, Annoyed, Guilty, Eye-opener - Adapted to Include Drugs) was incorporated into the AI prompts and scenarios. CAGE-AID is an extension of the original CAGE questionnaire designed to screen for both alcohol and other substance use disorders (Brown & Rounds, 1995). Including CAGE-AID in the prompt was important to ensure the scenarios addressed the screening aspect of SBIRT comprehensively for both alcohol and drug use.

Expert review and modified Delphi

To achieve consensus of expert opinion, we conducted a single-round modified Delphi survey, a method that uses a structured, anonymous group communication process to generate insights and build consensus among experts to evaluate the content quality of the AI-generated VR SBIRT simulation tool (Beiderbeck et al., 2021; Frallicciardi et al., 2016). This design was deliberately chosen as the study's objective was not to iterate to a perfect final product, but to conduct an initial content validation of the AI-generated scenarios. The goal was to determine if the AI content was a viable starting point, which was achieved by identifying clear patterns of strengths and weaknesses in the first round's ratings.

Table 1
Overview of SBIRT Scenarios.

Scenario	Focus	Key Learning Points
Scenario 1	Alcohol use	• Conduct the CAGE-AID screen. • Use motivational interviewing to explore readiness to change.
Scenario 2	Opioid use	• Screen for high-risk opioid use with CAGE-AID. • Discuss medication-assisted treatment (MAT) options and harm-reduction strategies.
Scenario 3	Pediatric substance use	• Perform age-appropriate substance-use assessment (CAGE-AID modified for adolescents). • Engage family/guardian when appropriate.

Note. CAGE-AID = CAGE Adapted to Include Drugs (Cut down, Annoyed, Guilty, Eye-opener); MAT = medication-assisted treatment. These scenarios represent common clinical situations requiring substance use screening and brief intervention. CAGE-AID is a validated four-question screening tool that can be used across different substance-use presentations.

To ensure a rigorous evaluation, we decided to assemble a panel of six highly qualified experts, all either certified healthcare simulation educators (CHSE) or CHSE advanced (CHSE-A), with a minimum of five years of experience. These specialists were tasked with conducting an blind review of the scenarios, utilizing the Simulation Scenario Evaluation Tool (SSET), a methodology developed by Hernandez et al. (2019). Five of the six content experts agreed to review the simulation scenarios, and the panel size (n = 5) was determined based on the principles of the Delphi method, which prioritizes the depth of panelist expertise over sample size. We employed the "key informant technique" a qualitative method for purposefully selecting individuals who, by their role and experience, possess a unique and comprehensive grasp of the subject matter (Marshall, 1996). Accordingly, our recruitment focused on a small, homogenous group of experts, all holding CHSE or CHSE-A certification with extensive experience to ensure the consensus achieved was based on high-level, specialized judgment. Each expert provided independent contractor certification and nondisclosure agreements prior to scenarios being sent out for review, and each received a small stipend for their time. To ensure the quality and relevance of these new scenarios, each expert was given all three scenario documents and independently evaluated each scenario via an online questionnaire. In Table 1, a description of the three AI-generated scenarios including their focus and key learning points is presented.

Based on the SSET framework, the evaluation form covered six key components of simulation design: learning objectives, clinical context, critical actions, patient trigger points/states, supporting materials/resources, and debriefing plan (Hernandez et al., 2019). These six elements were translated into a total of 20 specific items in the questionnaire, each rated on a 5-point scale (1 = Poor, 5 = Excellent). Each reviewer had a deadline of two weeks for completion of the reviews and an opportunity to provide written feedback on each scenario, after which all responses were collected for analysis.

Data analysis

After data collection from the SMEs, we separated quantitative and qualitative analysis to address different aims. For quantitative analysis, data were imported into the statistical program IBM SPSS V.30 and summarized with descriptive statistics (mean, standard

Table 2
Summary of Kappa Value.

Assessment Tool Category	Scenario 1	Scenario 2	Scenario 3	Overall κ	Agreement Level
Learning objectives	0.27	0.38	0.27	0.31	Fair
Clinical context/Scenario overview	0.70	0.37	0.58	0.55	Moderate
Critical actions	0.80	0.80	0.80	0.80	Substantial
Trigger points/patient states	0.33	0.40	0.35	0.36	Fair
Scenario materials and resources	0.20	0.40	0.40	0.33	Fair
Debriefing plan	0.70	0.38	0.00	0.36	Fair

Note. κ = Cohen's kappa coefficient. Agreement level interpretations: <0.20 = Poor; 0.21-0.40 = Fair; 0.41-0.60 = Moderate; 0.61-0.80 = Substantial; >0.80 = Almost Perfect (Landis & Koch, 1977).

Table 3
Intraclass Correlation Coefficient for Reliability of Raters Scores.

	Intraclass Correlation Coefficient						
	Intraclass Correlation[b]	95% Confidence Interval		F Test With True 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.0422[a]	−0.008	.380	2.284	4	76	0.068
Average Measures	.467[c]	−0.176	.925	2.284	4	76	0.068

Note. ICC = Intraclass Correlation Coefficient; CI = confidence interval; LL = lower limit; UL = upper limit. ICC values computed using a two-way mixed-effects model with people effects as random and measure effects as fixed. Single Measures ICC represents the reliability of individual ratings; Average Measures ICC represents the reliability of the mean of multiple ratings. Type A ICC (absolute agreement definition) was used.

deviation, minimum, and maximum) to examine overall scoring trends for the scenarios. Given the number of categories and raters, the inter-rater agreement was calculated using a kappa statistic (κ). A kappa statistic is considered more robust than simple percent agreement because it corrects for the possibility that agreement occurred by chance (McHugh, 2012). The specific statistic used was the free-marginal multirater kappa, and the analysis was conducted using the statistical software program R (V. 2024.09). We specifically used these statistics over the more common Fleiss kappa because it is known to be sensitive to rater bias which can paradoxically lower the kappa value even when agreement is high (a known issue called the "kappa paradox") (Randolph, 2010). We interpreted the kappa values using the classification guidelines from Landis and Koch (1977), where higher kappa indicates stronger consensus.

Inter-rater reliability among experts was also assessed by using a two-way mixed effects intraclass correlation coefficient (ICC) for absolute agreement across all ratings. Finally, to determine whether certain aspects of the scenarios were rated significantly higher or lower than others, we conducted a one-way ANOVA across the 20 items, followed by a posthoc Tukey test for pairwise comparisons. The qualitative comments provided by experts were reviewed separately, with recurring suggestions and criticisms noted for future revisions of both the scenarios and the evaluation instrument.

Results

All five expert reviewers completed the survey, and the VR scenarios received high overall ratings (M = 4.33 out of 5 on five-point scale, SD = 0.32, range = 3.89-4.68) indicating consistently high perceived scenario quality with minimal variability among raters. The reliability of the scored rater responses across the 20 proposed SSET assessment tool items was assessed via a two-way mixed ICC coefficient model for absolute agreement, and the value of 0.467 in Table 3 was determined moderately reliable. This ICC suggests that while there was some variability between individual expert's ratings, the average ratings could be considered for analysis purposes and supporting the use of mean scores for analysis.

For interpretation, Landis and Koch (1977) has proposed the following frequently used scale for kappa (κ) values: $\kappa < 0$ demonstrates poor agreement; 0.00 to 0.20 demonstrates slight agreement; 0.21 to 0.40 demonstrates fair agreement; 0.41 to 0.60 demonstrates moderate agreement; 0.61 to 0.80 demonstrates substantial agreement; and 0.81 to 1.00 demonstrates almost perfect agreement. For the critical action's element, a substantial consensus was achieved (κ = 0.80). Moderate agreement was observed for clinical context element (κ = 0.55). Lastly, fair agreement was observed for the elements: learning objectives, patient states, scenario materials and resources, and debriefing plan with kappa values ranging between 0.31 and 0.36 (See Table 2 for a summary of kappa values by element).

This divergence in ratings was confirmed by a one-way ANOVA, which indicated a statistically significant difference in mean ratings across the 20 items ($F(19,80) = 2.05$, $p = .014$; see Table 4 for ANOVA results). This finding demonstrates that the evaluation tool clearly delineated the scenarios strengths and weaknesses. For instance, the scenarios appeared strongest in covering the critical SBIRT actions required (received the highest scores and consensus), whereas they were relatively weaker in areas such as providing a comprehensive debriefing plan and including detailed supporting resources for facilitators (which had comparatively lower scores and agreement).

Overall, the participants gave positive feedback about the scenarios while providing useful recommendations to enhance their performance. To strengthen the learning experience, it was suggested that prebriefing sessions be incorporated to foster a culture of psychological safety, and that a standardized debriefing tool be utilized to facilitate more productive discussions. Furthermore, the importance of providing culturally relevant information was emphasized, enabling learners to adapt screening methods to diverse populations and integrate these assessments seamlessly into the CAGE-AID framework. Additional commentary highlighted the need for current references and proposed refinements to the narrative structure of the scenarios, specifically the interactions between

Table 4
One-Way Analysis of Variance for Score Differences.

Source	*SS*	*df*	*MS*	*F*	*p*
Between Groups	29.418	19	1.548	2.046	.014
Within Groups	60.537	80	0.757		
Total	89.956	99			

Note. SS = sum of squares; df = degrees of freedom; MS = mean square. The ANOVA revealed a significant difference in scores between groups, F (19,80) = 2.046, p = .014.

healthcare providers and patients, to create a more immersive and realistic learning environment.

Discussion

This study describes both the AI-assisted development of new SBIRT (Screening, Brief Intervention, and Referral to Treatment) focused VR simulation scenarios and an initial expert content validation of those scenarios. We modeled our single-round modified Delphi approach on the method used by Hernandez et al. (2019) to develop the original SSET for written simulation cases. In reviewing the literature, we found no prior existing instrument specifically designed to assess SBIRT focused VR scenarios, and our expert panel unanimously reported never having used a formal evaluation tool for this purpose. Thus, to our knowledge, this is the first structured, expert-consensus-derived use of the SSET to assess AI-generated SBIRT VR scenarios quality.

Our one-round modified Delphi was evaluated by five simulation educators and SBIRT practitioners. Based on Hernandez et al. (2019), we structured the evaluation around six core scenario elements: learning objectives, clinical context, critical actions, trigger points/patient states, supporting materials and resources, and debriefing plan, each rated on a 5-point scale. Although conducting a single Delphi round was a deliberate choice for initial validation rather than iterative refinement, this exercise still yielded clear patterns that both confirm strengths and reveal targets for improvement.

A key outcome of this research is the notable disparity in the level of agreement among raters when evaluating various components of simulated scenarios. The high agreement of consensus in "Critical Actions (κ = 0.80) shows that AI successfully retrieves and generates procedural knowledge which forms the fundamental framework of clinical interactions. In contrast, the AI system demonstrated fair agreement for aspects such as "Learning Objectives", "Patient States", "Supporting Materials", and "Debriefing Plan" (κ = 0.31-0.36), indicating limitations in its ability to grasp the nuances of simulation-based education and facilitation strategies. While AI can produce structured content at high speed, they lack the ability to develop complete learning experiences which requires human professionals to add educational value through collaborative work with subject matter experts.

In summary, for educators and simulation designers, an important takeaway from this project is that AI can create a basic outline of a realistic SBIRT scenario. However, it may miss important details that would make the simulation more effective for learners. Table 5 summarizes the key pros and cons we observed when leveraging ChatGPT for scenario development.

Limitations

A few limitations were identified with the initial development of the scenarios. First, this study focused on content validation of the three AI-generated VR SBIRT scenarios via expert review, rather than on the pilot outcomes of the over 500 students and practitioners who have since used the SBIRT VR platform at our institution. Second, the evaluation and pilot deployment occurred at a single institution, which may limit the generalizability of the findings. Third, this study represents a preliminary use of the SSET for VR scenarios, and the instrument is not yet fully validated for this specific modality. Despite these limitations, this project demonstrates the potential of the VR simulation platform to improve SBIRT training and highlights the ability of AI assistance to aid in building effective simulation scenarios.

Table 5
Pros and Cons of Using AI (ChatGPT) for Simulation Scenario Development.

Advantages of AI-Generated Scenario Development	Limitations of AI-Generated Scenario Development
Rapidly produces initial scenario drafts, greatly reducing development time.	May include inaccurate or outdated clinical information that requires correction by subject matter experts.
Provides a structured outline with dialogues and learning objectives already formulated.	May omit minute elements (e.g., a proper prebrief for psychological safety, or cultural considerations) unless explicitly prompted to include them.
Can generate diverse scenario ideas beyond the development team's immediate experience or expertise.	The generated content tends to be generic and might not align with specific curriculum objectives or institutional protocols without customization.
Allows easy iteration (developers can quickly adjust prompts to create variations of a case).	Requires thorough expert review and editing to ensure adherence, authenticity, and academic quality before implementation.

Note. AI = Artificial Intelligence. This table summarizes key considerations when using AI tools (e.g., ChatGPT) for clinical scenario development in simulation-based education. Both advantages and limitations should be carefully weighed when determining the appropriate role of AI in scenario creation workflows.

Funding

This work is supported by the Health Resources & Services Administration (HRSA) [HRSAT62HP49320, 2023].

Declaration of competing interest

None.

CRediT authorship contribution statement

Elizabeth Wells-Beede: Writing – original draft. **Lauren Thai:** Project administration. **Jinsil Hwaryoung Seo:** Conceptualization. **Mihir Sunil Godbole:** Conceptualization. **Jay Hareshbhai Patel:** Writing – review & editing, Methodology. **Cindy Weston:** Supervision. **Nicole Kroll:** Writing – review & editing, Project administration.

References

Beiderbeck, D., Frevel, N., von der Gracht, H. A., Schmidt, S. L., & Schweitzer, V. M. (2021). Preparing, conducting, and analyzing Delphi surveys: Cross-disciplinary practices, New directions, and Advancements. *MethodsX, 8*, Article 101401. https://doi.org/10.1016/j.mex.2021.101401.

Brown, R. L., & Rounds, L. A. (1995). Conjoint screening questionnaires for alcohol and other drug abuse: criterion validity in a primary care practice. *Wisconsin Medical Journal, 94*(3), 135–140. https://pubmed.ncbi.nlm.nih.gov/7778330/.

Fralliccardi, A., Vora, S., Bentley, S., Nadir, N.-A., Cassara, M., Hart, D., Park, C., Cheng, A., Aghera, A., Moadel, T., & Dobiesz, V. (2016). Development of an emergency medicine simulation fellowship consensus curriculum: Initiative of the society for academic emergency medicine simulation academy. *Academic Emergency Medicine, 23*(9), 1054–1060. https://doi.org/10.1111/acem.13019.

Hernandez, J., Fralliccardi, A., Nadir, N.-A., Gothard, M. D., & Ahmed, R. A. (2019). Development of a simulation scenario evaluation tool (SSET): Modified Delphi study. *BMJ Simulation and Technology Enhanced Learning, 6*(6), 344–350. https://doi.org/10.1136/bmjstel-2019-000521.

INACSL Standards Committee, Watts, P. I., McDermott, D. S., Alinier, G., Charnetski, M., Ludlow, J., Horsley, E., … Nawathe, P. A. (2021). Healthcare simulation standards of best practice simulation design. *Clinical Simulation in Nursing, 58*, 14–21. https://doi.org/10.1016/j.ecns.2021.08.009.

Landis, R. J., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310.

Marshall, M. (1996). The key informant technique. *Family Practice, 13*(1), 92–97. https://doi.org/10.1093/fampra/13.1.92.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemical Medical, 22*(3), 276. https://pmc.ncbi.nlm.nih.gov/articles/PMC3900052/.

OpenAI. (2024). ChatGPT. *ChatGPT; OpenAI*. https://chatgpt.com/.

Randolph, J. (2010). Free-marginal multirater kappa (multirater κfree): An alternative to fleiss fixed-marginal multirater kappa. *Advances in data analysis and classification*, 4. https://www.researchgate.net/publication/224890485_Free-Marginal_Multirater_Kappa_multirater_kfree_An_Alternative_to_Fleiss_Fixed-Marginal_Multirater_Kappa.

SAMHSA. (2024). *Screening, brief intervention, and referral to treatment (SBIRT)*. Samhsa.gov. https://www.samhsa.gov/substance-use/treatment/sbirt. Accessed December 22, 2025.

Yeh, S.-Y., Hassan, S., LaCaze, D., Weston, C. G., & Wells-Beede, E. (2024). Using Virtual Reality Simulation for Screening Brief Intervention and Referral to Treatment. *Journal of Nursing Education, 63*(7), 453–459. https://doi.org/10.3928/01484834-20240505-06.