Research Letter

Evaluating a ChatGPT Standardized Patient for Adolescent Contraceptive Counseling: Mixed Methods Pilot Study

Bianca Allison[1], MD, MPH; Anahita Gupta[2], BSPH; Emma Brockman[3], MEd; Cambray Smith[2,4], PhD; Lucy Elizabeth Burwell[5], MPH; Hailey Leiva[1], MSW; Carol Golin[6,7,8], MD; Samantha Schilling[1], MD, MSHP; Eric Zwemer[1], MD

[1]Department of Pediatrics, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[2]School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[3]School of Education, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[4]Department of Health Policy and Management, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[5]Department of Maternal and Child Health, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[6]Department of Health Behavior, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[7]Department of Internal Medicine, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[8]Center for AIDS Research, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

Corresponding Author:

Bianca Allison, MD, MPH
Department of Pediatrics
School of Medicine, University of North Carolina at Chapel Hill
231 MacNider Building, CB #7225
Chapel Hill, NC 27517
United States
Phone: 1 9849746669
Email: bianca.allison@unchealth.unc.edu

Abstract

This mixed methods pilot study, to develop and preliminarily evaluate a customized ChatGPT-based adolescent standardized patient to support clinicians in practicing person-centered contraceptive counseling, found that participants rated the AI simulation as highly feasible, acceptable, and appropriate, though automated grading showed limited agreement with human raters and requires further validation, supporting the simulation's perceived utility rather than reliable automated assessment.

JMIR Med Educ 2026;12:e99838; doi: 10.2196/99838

Keywords: standardized patient; pilot study; artificial intelligence; adolescent health; contraceptive counseling

Introduction

Person-centered contraceptive care supports adolescents' informed reproductive health decisions, yet many clinicians lack training and confidence in unbiased counseling [1]. Standardized patients offer a safe way to practice these skills, but are resource-intensive and rarely available to practicing clinicians [2]. AI may offer scalable practice and immediate feedback, with early evidence supporting its use as a standardized patient [3].

We conducted a mixed methods pilot to preliminarily evaluate a customized ChatGPT-based standardized patient for person-centered contraceptive care, assessing its feasibility, acceptability, appropriateness, and exploratory agreement between automated rubric scoring and human ratings.

Methods

Study Design

Grounded in Knowles' adult learning theory [4] (ie, adult self-direction and applied learning) and Kolb's experiential learning theory (ie, learning through practice and reflection) [5], we developed a ChatGPT standardized patient for

hands-on, problem-centered learning aligned with clinical practice. The simulated encounter and feedback supported experience and reflection, with subsequent clinical practice providing opportunities to apply learning.

We built a custom ChatGPT ("the GPT") using scenario goals and target competencies, selecting ChatGPT (OpenAI GPT-4.0, January 2025) for its customizability. The patient depicted a 16-year-old with heavy periods seeking confidential contraceptive counseling, expressing ambivalence and common concerns (eg, future fertility), requiring clinicians to explore preferences and support shared decision-making. Ten adolescent and sexual/reproductive health experts tested the GPT, informing refinements (eg, reduced patient transparency, added stage directions to convey uncertainty). The final instructions specified the patient persona, clinical background, encounter flow, feedback rules, and example responses (Multimedia Appendix 1).

We developed a 15-item rubric to capture observable counseling behaviors aligned with this scenario and curriculum (Multimedia Appendix 2), adapting 6 items from a contraceptive counseling observation checklist [6] and aligning remaining items with training goals [7]. Initial validity evidence included content validity (ie, expert review) and response process validity (ie, review of AI and human scoring). Written instructions directed the GPT to apply the 15-item rubric, display item-level scores, and provide narrative feedback. Two human raters scored the 10 expert transcripts, discussing each and achieving >90% agreement before scoring participant transcripts. Scoring discrepancies led to additional definitions for 6 rubric items in the GPT instructions (Multimedia Appendix 1).

Thirty-five adolescent-serving clinicians across North Carolina completed a continuing medical education–accredited training in person-centered contraceptive care in September and October 2024. In January 2025, participants were invited to complete a 30-minute Zoom session that included the GPT interaction, survey, and interview; 27 participated (77%).

After consenting, participants completed a 13-minute typed interaction with the GPT, followed by automated qualitative feedback (ie, 3 things they did well, 3 things to improve) and rubric-based scoring. The time limit approximated time spent counseling in clinical practice; completing all tasks was not required, and rubric scores were adjusted based on encounter completion. Participants then completed a 2-minute survey with demographics, and validated measures of feasibility, acceptability, and appropriateness (4 items each; 5-point Likert scale from 1 [completely disagree] to 5 [completely agree]) [8], and a 15-minute semistructured interview (Multimedia Appendix 3).

Descriptive statistics summarized demographics, and medians and ranges summarized implementation outcomes because distributions were skewed. Two human raters independently scored transcripts (91% agreement), resolving discrepancies by consensus. Human-AI scoring agreement was assessed using percent agreement and Cohen κ, which adjusts for chance agreement (1 indicates perfect agreement, 0 chance-level agreement, and negative values less agreement than expected by chance) [9]. Cohen κ is sensitive to the distribution of ratings and should be interpreted alongside percent agreement.

Qualitative interview data were analyzed using the rapid qualitative analysis [10] matrix method, organized by feasibility, acceptability, appropriateness, with double coding and discrepancies resolved through consensus. Quantitative and qualitative findings were integrated using a convergent mixed methods approach.

Ethical Considerations

The University of North Carolina School of Medicine Institutional Review Board approved this study (# 24-1207). Participants received a US $50 gift card. Interaction transcripts were deidentified and securely stored.

Results

Of the 27 participants, most were cisgender women (n=26, 96%), held an MD or DO degree (n=21, 78%), general pediatricians (n=25, 93%), and practiced in suburban settings (n=26, 96%) (Multimedia Appendix 4). Median ratings were 4.75 (range 3.25-5) out of 5 for acceptability, 4.75 (range 3.75-5) out of 5 for appropriateness, and 4.75 (range 3.25-5) out of 5 for feasibility. Across the 12 survey items, 96%-100% (n=26 and n=27, respectively) of the 27 participants agreed or completely agreed with each statement (Multimedia Appendix 5).

Human consensus-AI rubric grading was variable, with percent agreement ranging from 44.4% to 100% (Table 1). Five of the 9 items reported AI-human κ below 0.40 (indicating slight to fair agreement), whereas reported human-human κ ranged from 0.40 to 1.0 (indicating moderate to excellent agreement).

Table 1. Interrater reliability for human consensus and AI versus human consensus: Cohen κ and percent agreement by item.[a]

Rubric item	Human vs human: κ (percent agreement)	Human vs AI: κ (percent agreement)
1. Opened the conversation with a question (eg, "Would you like to talk about birth control today?")	1.0 (100)	N/A (96.3)
2. Explored patient's preferences and priorities for birth control method.	0.514 (74.1)	N/A (44.4)
3. Explored both patient's interest in pregnancy prevention (including past or future sexual activity) AND any interest in birth control beyond pregnancy prevention.	0.851 (92.6)	0.182 (63.0)
4. Explored patient's past experiences and knowledge of birth control methods.	0.786 (92.6)	N/A (77.8)

Rubric item	Human vs human: ϰ (percent agreement)	Human vs AI: ϰ (percent agreement)
5. Assessed where patient learned about birth control information (eg, online).	0.755 (92.6)	0.419 (85.2)
6. Concerns addressed respectfully (eg, "I hear you saying [blank] is a concern for you. Is it okay if I share with you what I've learned about that?")	N/A (100)	N/A (100)
7. Provided information on methods that align with patient preferences and goals	0.648 (96.0)	0.648 (96.0)
8. Addressed patient's myths or misperceptions about birth control by providing accurate information throughout the session.	0.706 (92.0)	0.288 (84.6)
9 Proactively counseled on potential side effects of selected method.	0.88 (94.4)	0.269 (68.4)
10. Assisted and supported birth control method selection including not using contraception if desired by patient.	0.409 (84.6)	0.632 (92.9)
11. Normalized ambivalence if it arose.	N/A (100)	N/A (85.7)
12. Used motivational interviewing techniques (open-ended questions, affirmations, reflection, and summarizing) during the session.	N/A (96.3)	N/A (96.3)
13. Normalized the adolescent's contraceptive journey (eg, "I think we can find a method that you like and works for you.")	0.752 (88.5)	0.196 (63.0)
14. Discussed following up to address questions, discuss side effects, or switch methods.	1.0 (100)	1.0 (100)
15. Mentioned or referred to using an adolescent-friendly decision aid or other resources for information (eg, social media accounts, websites)	0.595 (80.0)	0.359 (66.7)

[a]Some items have N/A (not applicable) for Cohen ϰ because of insufficient variability in ratings for that item; it does not indicate that the rubric item was inapplicable for the encounter. Percent agreement is shown for these items. ϰ adjusts for chance agreement; larger values indicate stronger agreement, with 1 indicating perfect agreement and 0 chance-level agreement.

Participants provided both positive and negative feedback about the ChatGPT standardized patient (Table 2). Specific benefits were comfort (n=13, 48%), realistic adolescent responses (n=11, 41%), and clear objective written feedback (n=10, 37%), realism of the clinical scenario (n=5, 19%), and a fun and novel experience (n=2, 7%). Limitations noted included absent nonverbal cues (n=15, 56%), which made it harder to interpret patient emotions; preference for verbal/audio interaction because typing slowed or/and disrupted flow (n=11, 41%), and insufficient time (n=10, 37%).

Table 2. Participant feedback on ChatGPT clinical encounter (N=27).

	Participants, n (%)	Representative quote[a]
Positive feedback		
More comfortable (eg, accessible, efficient, easier than clinical visit or standardized patient, less nervous, could think about best response)	13 (48)	Participant 6: "I actually really liked it too, because I think that sometimes, role playing face-to-face can be a little awkward. So, I actually found it to be a little bit easier to think about and type out my thoughts without having social interaction, for better or for worse."
		Participant 14: "I think it feels a lot less pressure just because I feel like when you're talking with the standardized patient, even though I know this is all in our head, but I feel so judged."
		Participant 25: "I think it's like it's easier because if, like you had been asking me, or like pretending to be the patient, I would have been like nervous and feel like I was acting, or whatever. So, being able to type, it made me a lot more comfortable, and I feel like I could get a lot more out of it practicing that way with, you know, ChatGPT versus like a person."
Similar to patient (eg, quality of responses, sounded like a real patient)	11 (41)	Participant 3: "I thought whoever wrote the responses for the teen did a good job of encapsulating how teens talk and ask questions and express concerns... It seemed fairly realistic in terms of the answers and questions that the teen had."
		Participant 16: "I felt like it was really accurate. I felt like I was actually talking to a teenager. They were very reactive to what I was saying and very easy to talk to and they kept up with everything that I was saying and answering all the specific questions."
Objective/helpful feedback (eg, rubric grading, feedback it gave)	10 (37)	Participant 8: "I thought it was kind of interesting and innovative and I like how it can give you feedback that's not necessarily biased, because sometimes I think the [human] actors, especially [if] they know you, they're not going to be mean to your face and like 'You could've said that better.' I guess I feel like maybe the ChatGPT is better at that."
		Participant 11: "The feedback portion I was not expecting and that was very impressive that it could pull everything out and tell me indicate what we had talked about and what had been what had been done and said."
Relevant clinical situation (eg, scenario seemed realistic)	5 (19)	Participant 1: "That's a scenario that I have encountered in clinic before. And it just gave me some things that I didn't like gave me a way to think about some things that I could address that I didn't address during the actual encounter."

	Participants, n (%)	Representative quote[a]
		Participant 16: "A lot of the concerns [from the ChatGPT] are definitely things that you hear mirrored in your own patients, specifically about their parents not trusting them and their parents not wanting to talk about it and just very, it was very realistic."
Fun (eg, enjoyed the simulation)	2 (7)	Participant 2: "Fun. I feel like such a millennial that I've never done something like that before. I liked it."
		Participant 10: "I think it was kind of fun to do."
Negative feedback		
Patient characteristics (eg, nonverbal cues, quality of responses, existing relationship)	15 (56)	Participant 13: "I feel like for most in the real world, most of my patients that I've built kind of a repertoire about with are a little bit more forthcoming usually when we're talking, because we've already built up that level of trust, and that rapport."
		Participant 14: "Nonverbals in a patient room are really helpful and so that was the only thing. Sometimes this is hard to... [read] it off a screen. Sometimes it's hard to see exactly what the patient is thinking and feeling about things, but there's only so much you can do when just reading it."
Didn't like typing/audio preferred	11 (41)	Participant 9: "The only thing that would have been nicer is to be able to talk to it. I feel like typing, it just takes you a little longer, and then I'm still trying to finish my thoughts and having to look back at what they just answered back. So anyway, it was just I felt like it was a little choppy in that sense."
		Participant 13: "I feel like conversations are just more fluid when you're able to use what you would actually say in clinic versus like what you would type. Because, of course, like, you don't really speak the way that you type so it'd be easier to get all those details, the things that you're looking for in the encounter, through natural speaking voice."
Time (eg, too short, different from constraints of clinic encounter)	10 (37)	Participant 18: "I was a little slow, so I didn't feel like I got to like totally flesh out the whole encounter, and address like all the needs that the patient brought up."
		Participant 26: "I think it's hard if you have a limited amount of time, because you're not going to… have time to complete the encounter so, you're going to feel like there are things that you didn't tie up."

[a]Quotes were edited to remove filler words.

Discussion

This pilot suggests that clinicians perceived a customized ChatGPT adolescent standardized patient as feasible, acceptable, and appropriate for practicing person-centered contraceptive care, supporting emerging evidence that AI standardized patients can be scalable tools for skills training [3]. While traditional approaches like human standardized patients and direct observation provide valuable human elements (eg, nonverbal cues), they are resource-intensive and often limited to structured training settings [2]. In contrast, AI-based simulations offer a complementary approach, enabling frequent, low-stakes, and self-directed practice beyond formal training.

Participants valued the qualitative feedback and overall experience. However, automated scoring requires refinement and independent validation before use as a standalone assessment of competence; human-AI scoring agreement was weaker than human-human reliability. Incorporating human-scored examples into grading instructions may improve scoring accuracy.

Limitations include a small, homogenous sample, convenience sampling, and facilitated sessions, which limits generalizability. The rubric was not formally validated, and the brief typed encounter may have restricted opportunities to sufficiently demonstrate rubric behaviors. Incorporating voice-based interactions and more diverse scenarios could address limitations in realism and broaden applicability. Feasibility and acceptability do not establish improvements in counseling skills or clinical outcomes, which should be assessed in a future study.

Acknowledgments

Thank you to the clinicians who participated in our study.

Generative AI Use

The authors declare that generative AI was used in the creation of this manuscript. Generative AI (ChatGPT) was used for proofreading, editing, or language polishing. The authors verified all outputs of the tool(s) used in the preparation of this manuscript and take full responsibility for the final content of the manuscript.

Funding

BA was partially supported by the Doris Duke Charitable Foundation [grant number 2020143], and the National Center for Advancing Translational Science (NCATS), National Institutes of Health [grant number 1K12TR004416-01]. CG was partially supported by the University of North Carolina at Chapel Hill Center for AIDS Research (P30 AI50410). CS is supported

through the National Institute of Child Health and Human Development at National Institutes of Health [1F30HD116454-01] and is a recipient of the Gertrude B. Elion Mentored Medical Student Research Award of Triangle Community Foundation. This manuscript's content is solely the responsibility of the authors and does not necessarily represent the official views of the funders. The funders were not involved in the collection, analysis, and interpretation of data; in authoring the report; and in submitting the article for publication.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: BA, CG, SS, EZ
Data curation: AG, EB, CS, LB
Formal analysis: BA, AG, EB
Funding acquisition: BA
Investigation: BA, AG, EB, CS, LB
Methodology: BA, EZ
Project administration: HL
Supervision: BA, CG, SS, EZ
Writing – original draft: BA, AG
Writing – review & editing: BA, EB, CS, LB, HL, CG, SS, EZ

Conflicts of Interest

None declared.

Multimedia Appendix 1

ChatGPT training document.
[DOCX File (Microsoft Word File), 36 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Person-centered contraceptive counseling standardized patient observation checklist.
[DOCX File (Microsoft Word File), 21 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Interview guide.
[DOCX File (Microsoft Word File), 21 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Participant characteristics.
[DOCX File (Microsoft Word File), 21 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Counts and percentages of each item within the measures acceptability, appropriateness, and feasibility.
[DOCX File (Microsoft Word File), 28 KB-Multimedia Appendix 5]

References

1. Thornton M, Mann ES, Bullington BW, Hartheimer J, Arora KS, Allison BA. Exploring adolescent-facing US clinicians' perceptions of their contraceptive counseling and use of shared decision-making: a qualitative study. Perspect Sex Reprod Health. Dec 2024;56(4):337-346. [doi: 10.1111/psrh.12283] [Medline: 39175306]
2. Elendu C, Amaechi DC, Okatta AU, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore). Jul 5, 2024;103(27):e38813. [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
3. Yamamoto A, Koda M, Ogawa H, et al. Enhancing medical interview skills through AI-simulated patient interactions: nonrandomized controlled trial. JMIR Med Educ. Sep 23, 2024;10:e58753. [doi: 10.2196/58753] [Medline: 39312284]
4. Knowles MS, Holton EF, Swanson RA. The Adult Learner: The Definitive Classic in Adult Education and Human Resource Development. 7th ed. Routledge; 2011.
5. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development. Prentice-Hall; 1984.
6. Contraceptive counseling observation tool. Cicatelli Associates Inc. URL: https://caiglobal.org/wp-content/uploads/2022/07/contraceptive-counseling-observation-tool.pdf [Accessed 2026-04-13]

7. Smith C, Leiva H, Patel J, et al. “I liked having my voice heard:” a feasible, rapid approach for codesigning a contraceptive counseling curriculum. Acad Med. Aug 1, 2026;101(8):931-942. [doi: 10.1093/acamed/wvag036] [Medline: 41691471]
8. Weiner BJ, Lewis CC, Stanick C, et al. Psychometric assessment of three newly developed implementation outcome measures. Implement Sci. Aug 29, 2017;12(1):108. [doi: 10.1186/s13012-017-0635-3] [Medline: 28851459]
9. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-282. [Medline: 23092060]
10. Kowalski CP, Nevedal AL, Finley EP, et al. Planning for and Assessing Rigor in Rapid Qualitative Analysis (PARRQA): a consensus-based framework for designing, conducting, and reporting. Implement Sci. Oct 11, 2024;19(1):71. [doi: 10.1186/s13012-024-01397-1] [Medline: 39394597]

Edited by Joshua Gentges; peer-reviewed by Deb Levine, Jeanelle Sheeder, Linda Reid-Chassiakos, Nicole Levitz; submitted 29.Apr.2026; initial comments to author 06.Sep.2026; final revised version received 03.Nov.2026; accepted 13.Nov.2026; published 23.Nov.2026

Please cite as:
Allison B, Gupta A, Brockman E, Smith C, Burwell LE, Leiva H, Golin C, Schilling S, Zwemer E
Evaluating a ChatGPT Standardized Patient for Adolescent Contraceptive Counseling: Mixed Methods Pilot Study
JMIR Med Educ 2026;12:e99838
URL: https://mededu.jmir.org/2026/1/e99838
doi: 10.2196/99838

© Bianca Allison, Anahita Gupta, Emma Brockman, Cambray Smith, Lucy Elizabeth Burwell, Hailey Leiva, Carol Golin, Samantha Schilling, Eric Zwemer. Originally published in JMIR Medical Education (https://mededu.jmir.org), 23.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.