TÜRK JİNEKOLOJİK ONKOLOJİ DERGİSİ
2025, VOL. 25 NO. 2, 24-33
https://doi.org/10.59215/trsgo.1762169

ORIGINAL ARTICLE/ORİJİNAL MAKALE

A New Ally in HPV and Cervical Cancer Screening: Age-Tailored Scenario-Based Analysis of ChatGPT

HPV ve Servikal Kanser Taramasında Yeni Müttefik: ChatGPT'nin Yaşa Duyarlı Senaryo Bazlı Analizi

Celal Akdemir[1]

[1] İzmir City Hospital, Department of Gynecologic Oncology, Istanbul, Türkiye

ABSTRACT

Introduction: Large language models such as ChatGPT are attracting increasing attention in health education. However, their reliability in providing age-sensitive, accurate, and guideline-compliant information about human papillomavirus (HPV) and cervical cancer screening has not been sufficiently investigated. This study aimed to evaluate the performance of ChatGPT-4 in terms of informativeness and guideline compliance when responding to HPV and cervical screening questions tailored to different age scenarios.

Methods: Thirty questions were developed based on three age scenarios (18, 30, and 45 years). Each question was submitted to the June 2025 version of ChatGPT-4. The responses were independently evaluated by a five-member panel consisting of three gynecologic oncology surgeons, one infectious diseases specialist, and one public health specialist. Evaluation was based on four criteria: scientific accuracy, clinical guideline compliance, comprehensibility (ease of understanding), and public health reliability. The term "comprehensibility" was used consistently throughout the study instead of "clarity". Each criterion was rated on a 5-point Likert scale.

Results: The overall mean score across all criteria was 4.19 ± 0.51. The highest mean score was for guideline consistency (4.22 ± 0.48), followed by public health reliability (4.20 ± 0.54), scientific accuracy (4.19 ± 0.50), and comprehensibility (4.16 ± 0.53). The 30-year-old scenario received the highest overall scores, particularly for scientific accuracy (4.34) and guideline consistency (4.26). The 18-year-old scenario scored highest in comprehensibility (4.28) but slightly lower in public health reliability (4.12). The 45-year-old scenario achieved the highest public health reliability score (4.32) but had marginally lower ratings for scientific accuracy (4.16) and comprehensibility (4.10). Expert comments highlighted ChatGPT's strengths in health communication and combating misinformation, while pointing out the lack of clinical details and explicit guideline references in some responses.

Conclusion: ChatGPT-4 appears to be an effective tool for promoting HPV vaccination and providing public health information, particularly in younger age groups. However, due to its limitations in clinical decision-making and guideline-based content, its use in patient education should be accompanied by expert oversight. Further research should encompass different model versions, additional evaluation metrics, and user perspectives.

Keywords: ChatGPT, Large Language Models, HPV, Cervical Cancer Screening, Age-Sensitive Information, Scenario-Based Analysis, Expert Evaluation, Health Education, Public Health Communication, Clinical Guideline Compliance

ÖZET

Giriş: ChatGPT gibi büyük dil modelleri sağlık eğitiminde giderek artan ilgi görmektedir. Ancak, insan papilloma virüsü (HPV) ve serviks kanseri taraması konusunda yaşa duyarlı, doğru ve kılavuza uyumlu bilgi sağlama güvenilirliği yeterince araştırılmamıştır. Bu çalışma, farklı yaş senaryolarına uyarlanmış HPV ve servikal tarama sorularına verdiği yanıtlarda ChatGPT-4'ün bilgilendiricilik ve kılavuza uyum performansını değerlendirmeyi amaçladı.

Yöntem: On sekiz, otuz ve kırk beş yaş olmak üzere üç yaş senaryosuna dayalı otuz soru hazırlandı. Her soru, ChatGPT-4'ün Haziran 2025 sürümüne sunuldu. Yanıtlar; üç jinekolojik onkoloji cerrahı, bir enfeksiyon hastalıkları uzmanı ve bir halk sağlığı uzmanından oluşan beş kişilik bir panel tarafından bağımsız olarak değerlendirildi. Değerlendirme; bilimsel doğruluk, klinik kılavuza uyum, anlaşılırlık (netlik ve kolay anlaşılabilirlik) ve halk sağlığı güvenilirliği olmak üzere dört ölçüte göre yapıldı. "Anlaşılırlık" kavramı çalışmada tutarlılık sağlamak amacıyla "comprehensibility" terimi ile ifade edildi. Her kriter 5 puanlık Likert ölçeği ile puanlandı.

Bulgular: Tüm kriterlerde genel ortalama puan 4,19 ± 0,51 idi. En yüksek ortalama puan kılavuza uyumda (4,22 ± 0,48) elde edildi; bunu halk sağlığı güvenilirliği (4,20 ± 0,54), bilimsel doğruluk (4,19 ± 0,50) ve anlaşılırlık (4,16 ± 0,53) izledi. Otuz yaş senaryosu, özellikle bilimsel doğruluk (4,34) ve kılavuza uyum (4,26) açısından en yüksek puanları aldı. On sekiz yaş senaryosu anlaşılırlıkta en yüksek puanı (4,28) elde etti ancak halk sağlığı güvenilirliği puanı biraz daha düşüktü (4,12). Kırk beş yaş senaryosu halk sağlığı güvenilirliğinde en yüksek puana ulaştı (4,32) ancak bilimsel doğruluk (4,16) ve anlaşılırlık (4,10) puanları biraz daha düşüktü. Uzman yorumları, ChatGPT'nin sağlık iletişimi ve yanlış bilgilendirmeyle mücadelede güçlü yönlerini vurgularken, bazı yanıtlarda klinik detayların ve açık kılavuz atıflarının eksikliğine dikkat çekti.

Sonuç: ChatGPT-4, özellikle genç yaş gruplarında HPV aşılamasını teşvik etme ve halk sağlığı bilgisi sağlama açısından etkili bir araç gibi görünmektedir. Ancak, klinik karar verme ve kılavuza dayalı içerikteki sınırlılıkları nedeniyle, hasta eğitiminde kullanımının uzman denetimiyle birlikte yürütülmesi önerilir. Gelecek araştırmalarda farklı model sürümleri, ek değerlendirme ölçütleri ve kullanıcı perspektifleri ele alınmalıdır.

Anahtar Kelimeler: ChatGPT, Büyük Dil Modelleri, HPV, Serviks Kanseri Taraması, Yaşa Duyarlı Bilgi, Senaryo Bazlı Analiz, Uzman Değerlendirmesi, Sağlık Eğitimi, Halk Sağlığı İletişimi, Klinik Kılavuz Uyumu

ARTICLE HISTORY
Recieved 10.08.2025
Accepted 25.08.2024

Correspondence: Celal Akdemir, İzmir City Hospital, Department of Gynecologic Oncology, Istanbul, Türkiye. E-mail: akdemircelal@gmail.com

Cite this Article: Akdemir C. A New Ally in HPV and Cervical Cancer Screening: Age-Tailored Scenario-Based Analysis of ChatGPT. The Turkish Journal of Gynecologic Oncology 2025;25(2):24-33.

Journal Websitesi: https://dergipark.org.tr/en/pub/trsgo **Publisher:** Cetus Publishing

INTRODUCTION

Human papillomavirus (HPV) is one of the most important causes of various gynecological and anogenital malignancies, including cervical, vulvar, vaginal, and anal cancers. However, one of the main factors limiting the effectiveness of current vaccination programs is the incomplete or incorrect public perception that HPV is associated only with cervical cancer. In a systematic review by Cangelosi et al. (1), it was emphasized that the success of HPV vaccination programs is adversely affected by knowledge gaps and misconceptions. This underscores the need for targeted, evidence-based educational interventions to increase community-based acceptance of preventive strategies in gynecologic oncology.

In recent years, large language models (LLMs) have emerged as potential tools in public health education. ChatGPT, one of the most well-known applications in this field, is increasingly used in areas such as patient education, clinical counseling, and raising public health awareness (2). However, Shen et al. (2) described such models as a “double-edged sword,” reporting that they may occasionally produce information of questionable accuracy, insufficient clinical details, and fabricated references (hallucinations).

Studies conducted in the context of HPV reveal both the potential and the limitations of ChatGPT. Patel et al. (3) evaluated responses to patient questions about HPV in terms of scientific accuracy, content completeness, and educational value; they found that only 45% of the responses were scientifically accurate and noted deficiencies particularly in guideline-based clinical management and follow-up recommendations. Nevertheless, the model’s ability to convey basic information in clear and understandable terms suggests it could serve as a complementary tool in public health communication. Similarly, Deiana et al. (4) highlighted that ChatGPT is generally accurate when addressing myths and misconceptions about vaccination, but its outputs may lack contextual nuance and consistent source validation, underscoring the need for professional oversight.

The use of ChatGPT in healthcare is not limited to public health education. Skryd and Lawrence (5) demonstrated that ChatGPT could serve as a potential educational tool for medical students and residents in clinical decision-making processes. In their study, the model was able to generate reasonable responses to complex clinical scenarios, but it was highlighted that expert supervision was essential for patient safety. Likewise, in a comprehensive systematic review, Li et al. (6) classified ChatGPT’s healthcare applications, identifying multiple potential areas such as patient education, clinical decision support, research processes, and public health communication, while also listing accuracy, source reliability, and contextual appropriateness as major limitations.

Studies investigating the direct impact of AI-based interventions on HPV vaccination rates are also available. Hou et al. (7) reported that a vaccine chatbot developed for parents significantly increased HPV vaccination rates among middle school-aged girls. This finding suggests that digital and AI-based tools could be effective in public health campaigns. However, ChatGPT’s performance may remain limited on HPV topics requiring detailed clinical knowledge. Bellamkonda et al. (8), in evaluating responses to frequently asked patient questions about HPV-positive oropharyngeal carcinoma, found that while the model performed well on some basic information, it lacked sufficient detail particularly in clinical management steps.

Message framing and persuasiveness are important factors influencing HPV vaccine acceptance. In a comparative study by Xia et al. (9), some pro-vaccine messages generated by ChatGPT were found to be more persuasive than those written by humans. This suggests that, when appropriate content and language tailored to the target audience are used, AI-based messages can be powerful tools in public health communication. These findings indicate that ChatGPT alone may not be sufficient, particularly in scenarios of vaccine hesitancy driven by cultural or belief-related factors.

In this context, evaluating ChatGPT's HPV-related outputs in age-specific scenarios based on criteria such as scientific accuracy, guideline compliance, comprehensibility, and public health reliability is important from both clinical and public health perspectives. This study aims to analyze ChatGPT-4's responses to HPV-related questions in 18-, 30-, and 45-year-old scenarios through a multi-criteria expert assessment.

MATERIALS AND METHODS

Study Design

This study was conducted using a content analysis methodology. The June 2025 version of the ChatGPT-4 model (ChatGPT Plus version) was used to address a total of 30 HPV-related questions structured according to three different age groups. Responses were recorded without any modifications. Since no data were collected from real individuals, ethics committee approval was not required.

Question Structure and Age Scenarios

Questions were developed according to three thematic age groups:

- 18 years: HPV vaccination, vaccine hesitancy, family pressure
- 30 years: Pap smear, transmission routes, partner trust
- 45 years: CIN classifications, colposcopy, follow-up recommendations

The content of the questions was based on patient knowledge gaps and common misconceptions about HPV identified in previous literature (4,6). Additionally, frequently asked questions from online patient forums, popular health platforms, and social media content were reviewed to reflect prevalent public knowledge gaps and misconceptions.

The full list of 30 questions and their sources is provided in Supplementary Table S1 to ensure transparency and reproducibility (Table S1).

Evaluation Panel and Criteria

Responses generated by ChatGPT were independently evaluated by a five-member panel consisting of three gynecologic oncology surgeons, one infectious diseases specialist, and one public health specialist. Each expert scored each response on a 5-point Likert scale across four domains:

1. Scientific accuracy
2. Guideline consistency (WHO, CDC, and up-to-date national/international guidelines)
3. Comprehensibility
4. Public health reliability

Although the panel included diverse subspecialists, the absence of patient representatives or primary care physicians may limit the assessment of comprehensibility from a broader audience perspective.

Data Analysis

For each question, the scores given by the five experts across the four criteria were compiled, resulting in a total of 600 evaluations. Data were analyzed by age group and criterion, and mean ± standard deviation values were calculated. Findings were supported with graphical and tabular presentations.

Reporting Principles

The methodological framework adhered to recommendations from prior systematic evaluations of ChatGPT in healthcare contexts (6). For consistency, the term "clarity" used in earlier drafts was standardized to "comprehensibility" to denote the ease of understanding of ChatGPT's responses.

RESULTS

A total of 600 individual ratings were collected from the five-member expert panel for 30 HPV-related questions, each evaluated across four predefined criteria: Scientific Accuracy, Guideline Consistency, Comprehensibility, and Public Health Reliability. Each criterion was scored on a 1–5 scale, with higher values indicating better performance.

The overall mean score for ChatGPT's responses across all criteria and experts was 4.19 ± 0.51, with 82% of all ratings in the 4–5 range, indicating generally accurate and educationally adequate content. No response received the lowest score of 1. Among the four criteria, Guideline Consistency achieved the highest mean score (4.22 ± 0.48), followed by Public Health Reliability (4.20 ± 0.54), Scientific Accuracy (4.19 ± 0.50), and Comprehensibility (4.16 ± 0.53). For consistency, the term "comprehensibility" is used throughout the manuscript to denote the ease of understanding of ChatGPT's responses. All numerical values correspond to those presented in Table 1.

Table 1. Overall Mean Scores

Criterion	Mean Score ± SD
Scientific Accuracy	4.19 ± 0.50
Guideline Consistency	4.22 ± 0.48
Comprehensibility	4.16 ± 0.53
Public Health Reliability	4.20 ± 0.54
*Overall Mean 4.19 ± 0.51	

When stratified by age scenario, the 30-year-old scenario received the highest overall scores, particularly for Scientific Accuracy (4.34) and Guideline Consistency (4.26). The 18-year-old scenario achieved the highest score in Comprehensibility (4.28) but slightly lower in Public Health Reliability (4.12). The 45-year-old scenario scored highest in Public Health Reliability (4.32) but marginally lower in Scientific Accuracy (4.16) and Comprehensibility (4.10) (Table 2). When tested with the Kruskal–Wallis test, no statistically significant differences were observed between the three age groups across any of the four evaluation criteria (all p > 0.05).

Table 2. Age-Specific Mean Scores

Age Group	Scientific Accuracy	Guideline Consistency	Comprehensibility	Public Health Reliability
18 years	4.08	4.16	4.28	4.12
30 years	4.34	4.26	4.10	4.16
45 years	4.16	4.24	4.10	4.32

* Values are presented as mean scores. Kruskal–Wallis test showed no statistically significant differences between the three age groups across any of the four evaluation criteria (all p > 0.05).

Mean scores per criterion according to expert specialty are shown in Table 3. Gynecologic oncology specialists consistently rated Guideline Consistency and Scientific Accuracy

Table 3. Expert-Based Average Scores per Evaluation Criterion

Criterion	Gynecologic Oncology 1	Gynecologic Oncology 2	Gynecologic Oncology 3	Infectious Diseases	Public Health
Scientific Accuracy	4.23	4.30	4.07	4.17	4.20
Guideline Consistency	4.47	4.27	4.07	4.30	4.00
Comprehensibility	4.07	4.10	4.10	4.13	4.40
Public Health Reliability	4.20	4.33	4.30	4.27	3.90

*Values are presented as mean scores. Inter-rater reliability analysis demonstrated low-to-fair agreement, with intraclass correlation coefficients (ICC) ranging from 0.14 to 0.22 across the four evaluation criteria.

Supplemantary Table S1.. List of HPV-Related Questions by Age Scenario and Their Sources

No	Age Scenario	Question	Source/Origin
1	18 years	Is it too late to get the HPV vaccine after the age of 18?	Guideline (WHO, CDC)
2	18 years	Does the vaccine cause infertility?	Common misconception / Social Media (YouTube comments, Twitter/X)
3	18 years	Can I get vaccinated without my parents' consent?	Patient forum (MedHelp, Reddit)
4	18 years	Is it only for women? Is it necessary for men as well?	Online Q&A (Quora, Google search results)
5	18 years	Can I have sexual intercourse immediately after vaccination?	Social Media (YouTube comments, Twitter/X)
6	18 years	Is it true that the HPV vaccine can be started at age 9?	Guideline (CDC, WHO)
7	18 years	I read on the internet that the HPV vaccine is dangerous; is that true?	Misconception / Social Media (YouTube comments, Twitter/X)
8	18 years	Do I need to have any other tests after the vaccination?	Guideline (CDC)
9	18 years	Where can I find the most reliable information about HPV?	Public health resources
10	18 years	Is the HPV vaccine covered by the government?	National health policy (country-specific)
1	30 years	My smear test is normal but I am HPV positive; what does this mean?	Common misconception / Social Media (YouTube comments, Twitter/X)
2	30 years	If I have HPV, does that mean my partner definitely cheated?	Patient forum / misconception (YouTube comments, Twitter/X)
3	30 years	Is it more dangerous if I have multiple HPV types?	Literature (HPV risk stratification studies)
4	30 years	If the screening test is positive, how often should I have follow-up?	Guideline (WHO, CDC)
5	30 years	Can HPV be transmitted from men to women?	Guideline (CDC)
6	30 years	Do condoms protect against HPV?	Literature + Guideline (WHO, CDC)
7	30 years	If I am HPV negative, does that mean I will never get cancer?	Patient forum (MedHelp, Reddit)
8	30 years	Can HPV be transmitted without sexual intercourse (e.g., swimming pool, toilet)?	Patient forum (MedHelp, Reddit)

9	30 years	Is a smear test the same as an HPV test?	Public health FAQ
10	30 years	Should I postpone pregnancy because I have HPV?	Social Media (YouTube comments, Twitter/X)
1	45 years	If I am HPV 16 positive, what is my cancer risk?	Literature (high-risk HPV studies)
2	45 years	What is the difference between CIN 1, 2, and 3?	Guideline (WHO classification)
3	45 years	Can HPV become active again years later?	Literature (HPV persistence/reactivation)
4	45 years	Is it mandatory to take a biopsy during colposcopy?	Patient forum (MedHelp, Reddit)
5	45 years	If HPV clears with immunity, does it leave any trace?	Literature (HPV natural history)
6	45 years	I am 45 and have never had a smear test; is it too late?	Guideline (WHO/CDC screening age limits)
7	45 years	If I am HPV positive, do I need a hysterectomy?	Misconception / Patient forum (MedHelp, Reddit)
8	45 years	Until what age should HPV screening be done?	Guideline (WHO/CDC, national protocols)
9	45 years	How often should I go for follow-up in HPV monitoring?	Guideline (WHO, CDC)
10	45 years	If HPV is contagious, how should my partner and I protect ourselves?	Guideline (WHO/CDC + public health FAQ)

the highest (mean range: 4.07–4.47). Public Health Reliability was rated highest by both gynecologic oncology and infectious diseases specialists (4.20–4.33), while the public health specialist assigned the highest Comprehensibility score (4.40). Variability between experts was generally low; however, the public health specialist rated Public Health Reliability slightly lower (3.90) compared with other panel members. The inter-rater reliability analysis demonstrated low-to-fair agreement between experts, with ICC values ranging from 0.14 to 0.22 across the four evaluation criteria.

Overall, experts agreed that ChatGPT's responses were well-aligned with current scientific guidelines and were presented in a clear, understandable manner. Noted limitations included the absence of explicit citations to guidelines, the lack of direct clinical directives, and occasional superficiality in complex clinical scenarios (e.g., CIN classification and colposcopy guidance). No response was deemed misleading, contradictory, or clinically unsafe.

DISCUSSION

The aim of this study was to investigate the prognostic value of adropin by determining the serum adropin level in EC patients. The prevalence of obesity and the associated cancer hazard has been ascending in the past several decades globally. In a study conducted in 2016, it was thought that approximately 2 billion adults and 340 million children worldwide have obesity problems (15). Given this increasing prevalence worldwide, the global obesity-related cancer burden is likely to increase in the future (16). Regarding gynecological cancers, an increase in BMI is associated with endometrial cancer rather than with ovarian cancer (17).

The presence of Diabetes Mellitus (DM) is associated with a worse prognosis in EC due to common risk factors such as obesity and age (18). Surveillance, Epidemiology and End Results data show variable increases in endometrial cancer incidence over time (14). The need for early diagnosis techniques with safe, fast and easy methods for clinical applicability is increasing.

Serum adropin levels are connected with coronary artery diseases, obesity, and obesity-related cancers (7,8). Meta-analysis results showed that serum adropin levels were significantly lower in patients with coronary artery disease, and then the possible relationship between serum adropin levels and the pathogenesis of coronary artery diseases was started to be investigated (7). The strongest information available today regarding the pathophysiology of coronary artery diseases is the coexistence of vascular inflammation, endothelial dysfunction and lipid metabolism disorder. Low adropin levels weaken endothelial protection and may cause atherosclerosis (19). Atherosclerosis is accelerated in both type 1 and type 2 DM. In addition, DM leads to decrease HDL, increase triglyceride (TG), and oxidative stress-induced endothelial dysfunction. The level of adropin increased with DM, which was accompanied by a reduction in fat accumulation, plasma TG, and inflammation (8).

Adropin has been defined as a possible regulatory hormone involved in the preservation of insulin sensitivity (20). In this study, Zang et al. proposed an optimal adropin cut-off value with a high sensitivity value (81.9%) to distinguish patients with type 2 diabetes from those without. Especially in obese patients, the adropin value was quite low. According to the results, adropin level was correlated negatively with age, parity, BMI, TG, DM, HT, insulin, HOMA-IR, and HbA1c, while positively with HDL (20). In another study, adropin value was significantly lower in the patients with cardiac syndrome (1.7 ± 0.8 ng/mL vs 3.4 ± 1.8 ng/mL; P <0.001), probably due to the difference in BMI values (28.1 ± 2.4 kg/m^2 vs 26.0 ± 3.7 kg/m^2 ; P < 0.001) (21).

Although our results showed lower adropin levels in EC group than control group, the difference was not significant. Contrary to our study, decreased plasma adropin concentrations were statistically significant in women diagnosed with EC (12). The difference in results may be related to the number of the subject.

A major limitation of this study is the relatively small sample size. Moreover, serum adropin levels were not assessed both at fasting and feeding conditions or after the surgery. In our study, no statistically significant difference in serum adropin level was found between EC and the control group. However, the difference was statistically significant between Type 1 and Type 2 EC. Adropin was statistically significant in type 2 EC in Roc analysis and logistic regression analysis. The reason for this result may be either type 2 EC cases are seen in older populations or randomly increased very-low-density lipoprotein (VLDL) and TG levels in our subjects. In our opinion, the statistical difference (p=0.01) is promising and this difference should be supported by higher sample size studies.

DISCUSSION

This study presents an original content analysis evaluating ChatGPT-4’s responses to HPV-related questions in terms of scientific accuracy, guideline compliance, comprehensibility, and public health reliability, based on age-specific scenarios. The findings indicate that the model performs particularly well in delivering age-tailored public health messages, but demonstrates limited guideline-based informational depth in complex clinical scenarios.

In our study, the highest mean scores were obtained in questions related to the 30-year-old group, particularly for scientific accuracy and guideline consistency. These scenarios often addressed issues of Pap smear interpretation, HPV transmission, and partner-related concerns, which may have allowed ChatGPT to generate more guideline-aligned and accurate responses.

The 18-year-old group achieved the highest comprehensibility score, reflecting the model's strength in delivering clear and accessible public health messages, especially regarding HPV vaccine safety, efficacy, and family-related hesitancy. Similarly, in the literature, Hou et al. (7) reported in a randomized controlled trial that digital information tools targeting specific groups increased HPV vaccination rates. This suggests that AI-based tools such as ChatGPT could be a strong complement to public health communication, particularly in younger populations.

Conversely, the 45-year-old group demonstrated comparatively lower scores in scientific accuracy and comprehensibility, underscoring the model's limitations in addressing more complex clinical topics. Evaluations by Patel et al. (3) and Bellamkonda et al. (8) likewise identified deficiencies in ChatGPT's recommendations for clinical management and follow-up. In particular, the absence of guideline-referenced information on CIN classifications, colposcopy referrals, and follow-up protocols for older patients parallels our findings.

Comprehensibility emerged as a relatively strong criterion across all age groups. This finding is consistent with the observations of Deiana et al. (4), who highlighted ChatGPT's ability to provide accurate yet oversight-dependent content, and Xia et al. (9), who showed that its pro-vaccination messages can even surpass human-written texts in persuasiveness. However, both Deiana et al. (4) and Passanante et al. (10) emphasized that expert supervision remains essential, particularly to ensure contextual appropriateness in clinical communication. Therefore, in gynecologic oncology practice, ChatGPT should function only as a supportive tool rather than an autonomous source of patient education.

The literature on the use of LLMs in gynecologic oncology-specific domains is limited. Kuerbanjiang et al. (11) evaluated LLM performance in cervical cancer management and found that while basic information delivery was adequate, clinical decision support was limited. Similarly, Angyal et al. highlighted the potential of LLMs in cervical cancer screening education. Together with our findings, these studies indicate that AI tools can be powerful for education and awareness-raising purposes but should be used cautiously for clinical decision support.

One of the strengths of our study is the multidisciplinary nature of the evaluation panel. Notably, the public health specialist assigned a slightly lower score for Public Health Reliability compared with other experts, possibly reflecting the application of stricter criteria for population-level reliability and evidence integration. Perspectives from specialists in obstetrics and gynecology, infectious diseases, and public health provided a multidimensional analysis of ChatGPT's responses. However, certain limitations should be acknowledged. First, the evaluation was limited to GPT-4; no comparisons were made with GPT-3.5 or future models. Second, the analysis was based solely on expert assessments, without incorporating patient or public perspectives. Furthermore, the reliability and traceability of references in the responses were not systematically evaluated. This omission is relevant given the well-documented "hallucination" phenomenon in large language models, which refers to the generation of inaccurate or fabricated information despite confident presentation (2,4,12,13). Previous studies have highlighted that such inaccuracies may undermine clinical reliability, particularly when AI outputs are used without expert oversight, underscoring the importance of systematic reference verification in future research. These limitations, including the lack of patient or lay perspectives, the restriction to GPT-4, and the absence of systematic reference verification, are of critical importance as they directly affect the reproducibility and generalizability of our findings

Overall, our study shows that ChatGPT has high potential for age-specific preventive health messaging on HPV, especially in younger age groups, but its limitations in producing guideline-based content in gynecologic oncology should be noted. In clinical practice, ChatGPT should be used under professional supervision for educational and awareness purposes, and positioned as a complementary rather than a directive tool in clinical decision-making. Future research should include comparative analyses of different LLM versions, evaluations incorporating patient and public perspectives, and applications in other areas of gynecologic oncology. These approaches will contribute to enhancing the reliability and effectiveness of AI-based educational tools, while addressing current limitations that constrain reproducibility and generalizability.

ACKNOWLEDGEMENT

Conflicts of interest

The authors have no conflicts of interest.

Funding

No funds were received in support of this work.

Ethical approval

This study did not involve human participants or patient data; therefore, ethical approval was not required.

REFERENCES

1. Cangelosi G, Sacchini F, Mancin S, Petrelli F, Amendola A, Fappani C, Sguanci M, Morales Palomares S, Gravante F, Caggianelli G. Papillomavirus vaccination programs and knowledge gaps as barriers to implementation: a systematic review. Vaccines (Basel). 2025 Apr 25;13(5):460.
2. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, Moy L. ChatGPT and other large language models are double-edged swords. Radiology. 2023 Apr;307(2):e230163.
3. Patel TA, Michaelson G, Morton Z, Harris A, Smith B, Bourguillon R, Wu E, Eguia A, Maxwell JH. Use of ChatGPT for patient education involving HPV-associated oropharyngeal cancer. Am J Otolaryngol. 2025 Jul-Aug;46(4):104642.
4. Deiana G, Dettori M, Arghittu A, Azara A, Gabutti G, Castiglia P. Artificial intelligence and public health: evaluating ChatGPT responses to vaccination myths and misconceptions. Vaccines (Basel). 2023 Jul 7;11(7):1217.
5. Skryd A, Lawrence K. ChatGPT as a tool for medical education and clinical decision-making on the wards: case study. JMIR Form Res. 2024 May 8;8:e51346.
6. Li J, Dada A, Puladi B, Kleesiek J, Egger J. ChatGPT in healthcare: a taxonomy and systematic review. Comput Methods Programs Biomed. 2024 Mar;245:108013.
7. Hou Z, Wu Z, Qu Z, Gong L, Peng H, Jit M, Larson HJ, Wu JT, Lin L. A vaccine chatbot intervention for parents to improve HPV vaccination uptake among middle school girls: a cluster randomized trial. Nat Med. 2025 Jun;31(6):1855-1862.
8. Bellamkonda N, Farlow JL, Haring CT, Sim MW, Seim NB, Cannon RB, Monroe MM, Agrawal A, Rocco JW, McCrary HC. Evaluating the accuracy of ChatGPT in common patient questions regarding HPV+ oropharyngeal carcinoma. Ann Otol Rhinol Laryngol. 2024 Sep;133(9):814-819.
9. Xia D, Song M, Zhu T. A comparison of the persuasiveness of human and ChatGPT generated pro-vaccine messages for HPV. Front Public Health. 2025 Jan 16;12:1515871.
10. Passanante A, Pertwee E, Lin L, et al. Conversational AI and vaccine communication: systematic review of the evidence. J Med Internet Res. 2023;25:e42758.
11. Kuerbanjiang W, Peng S, Jiamaliding Y, Yi Y. Performance evaluation of large language models in cervical cancer management based on a standardized questionnaire: comparative study. J Med Internet Res. 2025;27:e63626.

12. Cosma C, Radi A, Cattano R, et al. Potential role of ChatGPT in simplifying and improving informed consent forms for vaccination: a pilot study conducted in Italy. BMJ Health Care Inform. 2025;32(1):e101248.

13. Cheng K, Li Z, He Y, et al. Potential use of artificial intelligence in infectious disease: take ChatGPT as an example. Ann Biomed Eng. 2023;51(6):1130-1135.