

Age-Stratified Evaluation of ChatGPT Recommendations for Abnormal Uterine Bleeding: An Expert-Based Assessment of Guideline Adherence and Clinical Appropriateness

İ Süleyman Özen¹, İ Eda Güner Özen², İ Muzaffer Sancı¹

¹ Department of Gynecologic Oncology, İzmir City Hospital, İzmir, Turkey

² Department of Obstetrics and Gynecology, İzmir City Hospital, İzmir, Turkey

*Correspondence: Süleyman Özen; drsuleymanozen@gmail.com

Cite as: Özen S, Güner Özen E, Sancı M. Age-Stratified Evaluation of ChatGPT Recommendations for Abnormal Uterine Bleeding: An Expert-Based Assessment of Guideline Adherence and Clinical Appropriateness. J Cukurova Anesth Surg. 2026;9(3):733-745. doi:10.36516/jocass.1988506

Received: 06.07.2026, **Accepted:** 17.09.2026, **Published:** 30.09.2026, CC BY-NC 4.0

Abstract

Background: Abnormal uterine bleeding (AUB) requires age-sensitive evaluation because endometrial hyperplasia or malignancy risk increases with age and after menopause. Large language models are increasingly used for health information, but their performance in age-dependent gynecologic triage remains unclear. **Objective:** To evaluate the clinical appropriateness, guideline adherence, safety, completeness, endometrial sampling accuracy, malignancy-risk recognition, and urgency classification of ChatGPT responses to standardized AUB and postmenopausal bleeding scenarios. **Methods:** Sixty investigator-generated scenarios were reviewed by two gynecologic oncologists and stratified into five groups: adolescents, reproductive-age women, late reproductive-age women, premenopausal/perimenopausal women aged 45–55 years, and women with postmenopausal bleeding. Each scenario was submitted to ChatGPT using a standardized prompt. Responses were independently evaluated by two gynecologic oncologists across four domains: clinical appropriateness, guideline adherence, patient safety, and completeness. **Results:** The overall expert-rated appropriateness score was 3.48 ± 0.66 , and 21 of 60 responses (35.0%) were appropriate. Mean overall scores decreased from 4.10 ± 0.41 in adolescents and 3.92 ± 0.33 in reproductive-age women to 2.98 ± 0.49 in women aged 45–55 years and 2.82 ± 0.51 in postmenopausal bleeding scenarios ($p < 0.001$). Endometrial sampling recommendations were correct in 72 of 120 reviewer-level assessments (60.0%). High-risk or malignancy-related features were recognized in 19 of 40 relevant assessments (47.5%). Inter-rater reliability for overall score was excellent (ICC = 0.957). **Conclusion:** ChatGPT performed better in adolescent and reproductive-age AUB scenarios but showed lower appropriateness in age groups with higher endometrial pathology risk. Clinician oversight remains necessary when AI-generated recommendations may influence AUB triage or endometrial cancer risk assessment.

Keywords: abnormal uterine bleeding; artificial intelligence; ChatGPT; endometrial sampling; large language model; postmenopausal bleeding

1. Introduction

Abnormal uterine bleeding (AUB) is among the most common gynecologic complaints encountered throughout a woman's lifespan and represents a major reason for referral to gynecologic services worldwide. Beyond its high prevalence, AUB imposes a substantial burden on affected individuals through its impact on physical health, emotional well-being, social functioning, work productivity, and healthcare utilization¹. Because the underlying causes of AUB encompass a broad spectrum of physiologic, benign, premalignant, and malignant conditions, accurate evaluation is essential to ensure appropriate management and timely identification of clinically significant pathology.

To facilitate a standardized approach to diagnosis and research, the International Federation of Gynecology and Obstetrics (FIGO) introduced the PALM–COEIN classification system, which categorizes AUB according to structural causes—polyp, adenomyosis, leiomyoma, and malignancy/hyperplasia—and non-structural causes—coagulopathy, ovulatory dysfunction, endometrial disorders, iatrogenic causes, and not otherwise classified². The classification was subsequently updated and refined and has become an international standard for the evaluation and reporting of AUB^{3,4}.

The clinical significance of AUB varies substantially across age groups. In adolescents and younger reproductive-age women, abnormal bleeding is commonly associated with anovulation, coagulation disorders, and transient hormonal disturbances. In contrast, increasing age is associated with a progressively higher prevalence of endometrial hyperplasia, atypical hyperplasia, and endometrial carcinoma^{4,5}. Consequently, age is a major determinant of the diagnostic approach and influences both the intensity of investigation and the threshold for invasive assessment.

Endometrial sampling remains a cornerstone of evaluation in women presenting with AUB who are at increased risk of endometrial pathology. Current recommendations advocate routine endometrial assessment in women aged 45 years and older with abnormal uterine bleeding and in younger women with persistent symptoms, risk factors for unopposed estrogen exposure, or failure of medical management⁵. Likewise, postmenopausal bleeding requires prompt investigation because it may represent the earliest manifestation of endometrial malignancy^{6,7}. Although many cases ultimately prove to have benign causes, postmenopausal bleeding remains one of the most important presenting symptoms of endometrial cancer.

The importance of correctly identifying patients who require endometrial assessment is underscored by the burden of endometrial cancer, which is the most common gynecologic malignancy in many developed countries⁷. Delayed recognition of malignancy risk factors may postpone diagnosis and treatment, emphasizing the importance of guideline-based triage and management strategies.

In parallel with advances in clinical medicine, large language models (LLMs) have rapidly emerged as widely accessible sources of health information. Since the public introduction of ChatGPT, generative artificial intelligence systems have increasingly been used for medical information retrieval, symptom interpretation, and clinical decision support. Early studies demonstrated that LLMs can generate clinically coherent responses and perform well on standardized medical examinations^{8,9}.

Within women's health, obstetrics, gynecology, and gynecologic oncology, a growing body of literature has explored the reliability and clinical utility of LLM-generated recommendations. While several studies have reported promising performance in educational and informational tasks, others have highlighted limitations related to guideline adherence, contextual reasoning, recognition of high-risk conditions, and consistency of recommendations^{10–16}. Particular concern has focused on hallucinations, omission of critical diagnostic steps, failure to recognize red-flag symptoms, and provision of falsely reassuring recommendations in situations requiring urgent medical attention^{13–16}.

Despite growing interest in the application of LLMs to women's health, evidence regarding their performance in the evaluation of abnormal uterine bleeding remains limited. Existing studies have primarily focused on examination-style questions, patient education materials, general gynecologic counseling, or oncology-related topics. Whether the quality of LLM-generated recommendations varies according to patient age in AUB assessment—and particularly whether these systems appropriately identify indications for endometrial sampling and malignancy evaluation—has not been systematically investigated.

Given the central role of age in determining the diagnostic approach to AUB and the increasing reliance on generative artificial intelligence for health-related information, evaluating age-specific LLM performance has important implications for patient safety. Therefore, the aim of this study was to evaluate the clinical appropriateness, guideline adherence, safety, completeness, endometrial sampling accuracy, malignancy-risk recognition, and urgency classification of ChatGPT-generated responses to standardized AUB and postmenopausal bleeding scenarios across different age groups. We hypothesized that performance would decline in clinical scenarios where age-related cancer risk becomes a major determinant of management.

2. Materials and Methods

2.1. Study Design

This cross-sectional, scenario-based expert evaluation study was designed to assess the clinical appropriateness, guideline adherence, safety, and completeness of ChatGPT-generated recommendations for abnormal uterine bleeding across different age groups. This study involved only investigator-generated hypothetical clinical scenarios and expert evaluation of artificial intelligence-generated responses and did not include human participants, real patient data, biological materials, clinical records, images, institutional health data, or identifiable personal information. The study protocol was approved by the İzmir City Hospital Ethics Committee on June 17, 2026 (decision no: 2026/459).

The primary objective was to evaluate the overall expert-rated appropriateness of ChatGPT responses to age-stratified abnormal uterine bleeding scenarios. Secondary objectives included the assessment of clinical appropriateness, guideline adherence, patient safety, completeness, endometrial sampling accuracy, urgency classification, recognition of high-risk or malignancy-related features, harmful or falsely reassuring recommendations, and inter-rater agreement between two gynecologic oncology specialists.

2.2. Development of Clinical Scenarios

A total of 60 standardized clinical scenarios were developed by the investigators to represent common and clinically relevant presentations of abnormal uterine bleeding and postmenopausal bleeding encountered in gynecologic practice. Scenario development was guided by contemporary international recommendations and the FIGO PALM–COEIN classification framework².

The predefined reference standard for guideline adherence was based on the FIGO PALM–COEIN classification system and its updated terminology for AUB classification and reporting^{2–4}, contemporary recommendations regarding endometrial sampling in premenopausal, perimenopausal, and postmenopausal abnormal uterine bleeding⁵, and ACOG Committee Opinion No. 734 for the evaluation of postmenopausal bleeding and endometrial assessment⁶.

The scenarios were designed to cover a broad clinical spectrum, including ovulatory dysfunction or anovulatory bleeding, endometrial polyps, leiomyomas, high-risk abnormal uterine bleeding, suspected endometrial hyperplasia or malignancy, acute severe bleeding, and postmenopausal bleeding. To evaluate the influence of age on the quality

of AI-generated recommendations, scenarios were stratified into five predefined age groups: adolescents aged 13–18 years, reproductive-age women aged 19–39 years, late reproductive-age women aged 40–44 years, premenopausal or perimenopausal women aged 45–55 years, and women with postmenopausal bleeding.

Each age group included 12 unique scenarios, resulting in a total of 60 scenarios. Each scenario included relevant clinical information such as patient age, bleeding pattern, menstrual or menopausal status, medical history, risk factors, physical examination findings, laboratory parameters, and ultrasonographic findings when applicable.

Before AI assessment, all scenarios were independently reviewed by both gynecologic oncologists to ensure clinical realism, internal consistency, and alignment with contemporary AUB and postmenopausal bleeding guidance. Any discrepancies regarding scenario plausibility or expected reference-standard decisions were resolved by consensus before the scenarios were submitted to ChatGPT.

Before artificial intelligence assessment, expected reference-standard decisions were predefined for each scenario by the investigator team according to the guideline-based reference standard described above. These included the expected need for endometrial sampling, urgency level, recognition of high-risk or malignancy-related features, and appropriateness of diagnostic and management recommendations. Reference-standard decisions were not modified after ChatGPT responses were generated.

Ordinal domain scores were summarized at the scenario level after averaging the two reviewers' ratings. Because each of the 60 scenarios was independently evaluated by two gynecologic oncologists, binary outcome analyses were performed at the reviewer-assessment level, yielding 120 evaluations. Recognition of high-risk or malignancy-related features was evaluated only in scenarios predefined as containing such features; this analysis included 20 scenarios and 40 reviewer-level assessments.

2.3. Artificial Intelligence Assessment

ChatGPT-5.5 Thinking (OpenAI, San Francisco, CA, USA) was accessed through the ChatGPT web interface for all AI-generated responses. Each clinical scenario was independently submitted to ChatGPT using an identical standardized prompt to minimize prompt-related variability. The prompt requested a structured gynecologic assessment including differential diagnosis, recommended investigations, need for endometrial sampling, initial management, urgency level, and identification of red-flag features requiring prompt or emergency evaluation.

The standardized prompt was as follows:

"You are a board-certified gynecologist. Evaluate the following patient with abnormal uterine bleeding. Provide:

- 1. Differential diagnosis*
- 2. Recommended investigations*
- 3. Need for endometrial sampling*
- 4. Initial management*
- 5. Urgency level*

Base your answer on current international gynecology guidelines.

Clinical scenario: [scenario text]"

Each scenario was entered separately. No feedback, clarification, prompt modification, follow-up prompting, or iterative correction was applied after the initial response. ChatGPT-generated responses were recorded verbatim and anonymized before expert evaluation. Access platform and query dates were retained in the study log for reproducibility.

2.4. Expert Evaluation

Two board-certified gynecologic oncologists independently evaluated all ChatGPT-generated responses using a structured assessment form. The reviewers were blinded to each other's assessments throughout the evaluation process. Because patient age was an essential component of each clinical scenario, blinding to age group was not feasible.

Each response was assessed across four predefined domains: clinical appropriateness, guideline adherence, patient safety, and completeness. Clinical appropriateness reflected whether the proposed diagnostic and management approach was clinically reasonable for the given scenario. Guideline adherence reflected consistency with the predefined reference standard based on FIGO AUB classification and terminology, contemporary endometrial sampling recommendations, and ACOG guidance for postmenopausal bleeding²⁻⁶. Patient safety reflected avoidance of potentially harmful, delayed, or falsely reassuring recommendations. Completeness was defined as the extent to which the response addressed all clinically essential components expected for the scenario, including relevant differential diagnoses, appropriate initial investigations, indication or non-indication for endometrial sampling, initial management, urgency classification, and red-flag features.

Each domain was scored using a five-point Likert scale, where 1 indicated an inappropriate or potentially harmful response, 2 indicated major deficiencies, 3 indicated partial appropriateness, 4 indicated an appropriate response with minor omissions, and 5 indicated a fully appropriate and guideline-concordant response. The overall scenario-level score was calculated as the mean of the four domain scores averaged across the two reviewers.

In addition to ordinal scoring, the reviewers assessed predefined binary outcomes, including correct recommendation for endometrial sampling, correct urgency classification, appropriate recognition of high-risk or malignancy-related features, and presence of potentially harmful or falsely reassuring recommendations. Harmful or falsely reassuring recommendations were defined as responses that could delay appropriate clinical evaluation, omit a necessary biopsy or urgent referral, underestimate malignancy risk, or provide reassurance despite the presence of high-risk features.

For overall interpretation, responses were categorized as appropriate when the overall mean score was ≥ 4.0 , partially appropriate when the score was 3.0–3.99, and inappropriate when the score was < 3.0 .

2.5. Outcome Measures

The primary outcome was the overall expert-rated appropriateness of ChatGPT responses. Secondary outcomes included clinical appropriateness score, guideline adherence score, patient safety score, completeness score, correctness of endometrial sampling recommendations, correctness of urgency assessment, recognition of high-risk or malignancy-related features, presence of harmful or falsely reassuring recommendations, and inter-rater agreement.

High-risk or malignancy-related feature recognition was evaluated only in scenarios that were predefined as containing high-risk or malignancy-related features. This analysis included 20 scenarios, resulting in 40 reviewer-level assessments because each scenario was independently evaluated by two gynecologic oncology specialists.

2.6. Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics version 26.0 software (IBM Corp., Armonk, NY, USA). Continuous variables were summarized as mean $\pm$ standard deviation or median with interquartile range according to data distribution. Categorical variables were presented as frequencies and percentages.

Differences in expert-rated scores across age groups were evaluated using the Kruskal–Wallis test. When statistically significant differences were detected, pairwise comparisons were performed using Dunn’s post hoc procedure with Bonferroni correction. The association between increasing age stratum and overall appropriateness score was assessed using Spearman correlation analysis. Ordinal logistic regression was used as an exploratory analysis to estimate the relationship between age stratum and appropriateness category.

Inter-rater reliability for ordinal scores was assessed using the intraclass correlation coefficient. Agreement for binary outcomes was evaluated using Cohen’s kappa coefficient. Intraclass correlation coefficient values above 0.75 were interpreted as indicating good reliability, and values above 0.90 were interpreted as indicating excellent reliability.

A two-sided *p* value of <0.05 was considered statistically significant.

3. Results

3.1. Scenario Characteristics

A total of 60 clinical scenarios were included in the analysis, with 12 scenarios allocated to each of the five predefined age strata. The median age represented across all clinical scenarios was 42.5 years, with an interquartile range of 30.5–52.2 years. Median ages were 16.5 years in adolescents, 34.5 years in reproductive-age women, 42.5 years in late reproductive-age women, 49.5 years in premenopausal or perimenopausal women aged 45–55 years, and 65.5 years in women with postmenopausal bleeding.

The scenarios covered clinically relevant AUB and postmenopausal bleeding presentations, including ovulatory dysfunction or anovulation, endometrial polyp, leiomyoma, hyperplasia or high-risk AUB, suspected malignancy, acute severe bleeding, and low-risk postmenopausal bleeding differentials. Expected endometrial sampling was classified as required in 28 scenarios, conditional or individualized in 22 scenarios, and not routinely indicated in 10 scenarios. Expected urgency was classified as routine outpatient evaluation in 6 scenarios, prompt or urgent outpatient evaluation in 44 scenarios, and emergency or emergent evaluation in 10 scenarios. The distribution of scenario characteristics according to age stratum is summarized in Table 1.

Table 1. Scenario characteristics by age stratum

Characteristic	Total (n=60)	Adolescents (13–18 y) (n=12)	Reproductive age (19–39 y) (n=12)	Late reproductive age (40–44 y) (n=12)	45–55 y pre/perimenopausal (n=12)	Postmenopausal bleeding (n=12)
Age, years, median (IQR)	42.5 (30.5–52.2)	16.5 (15.0–17.2)	34.5 (30.5–37.2)	42.5 (41.0–44.0)	49.5 (46.8–52.2)	65.5 (61.0–69.8)
Scenario domain						
Ovulatory dysfunction/anovulation	8 (13.3)	2 (16.7)	2 (16.7)	2 (16.7)	2 (16.7)	0 (0.0)
Polyp	10 (16.7)	2 (16.7)	2 (16.7)	2 (16.7)	2 (16.7)	2 (16.7)
Leiomyoma	10 (16.7)	2 (16.7)	2 (16.7)	2 (16.7)	2 (16.7)	2 (16.7)
Hyperplasia/high-risk AUB	10 (16.7)	2 (16.7)	2 (16.7)	2 (16.7)	2 (16.7)	2 (16.7)
Malignancy suspicion	10 (16.7)	2 (16.7)	2 (16.7)	2 (16.7)	2 (16.7)	2 (16.7)
Acute severe bleeding	10 (16.7)	2 (16.7)	2 (16.7)	2 (16.7)	2 (16.7)	2 (16.7)
Low-risk PMB differential	2 (3.3)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	2 (16.7)
Expected endometrial sampling						
Required	28 (46.7)	0 (0.0)	3 (25.0)	3 (25.0)	12 (100.0)	10 (83.3)

Characteristic	Total (n=60)	Adolescents (13–18 y) (n=12)	Reproductive age (19–39 y) (n=12)	Late reproduc- tive age (40–44 y) (n=12)	45–55 y pre/perimeno- pausal (n=12)	Postmenopau- sal bleeding (n=12)
Conditional/individualized	22 (36.7)	5 (41.7)	6 (50.0)	9 (75.0)	0 (0.0)	2 (16.7)
Not routinely indicated	10 (16.7)	7 (58.3)	3 (25.0)	0 (0.0)	0 (0.0)	0 (0.0)
Expected urgency level						
Routine outpatient	6 (10.0)	2 (16.7)	3 (25.0)	1 (8.3)	0 (0.0)	0 (0.0)
Prompt/urgent outpatient	44 (73.3)	8 (66.7)	7 (58.3)	9 (75.0)	10 (83.3)	10 (83.3)
Emergency/emergent	10 (16.7)	2 (16.7)	2 (16.7)	2 (16.7)	2 (16.7)	2 (16.7)

Abbreviations: AUB, abnormal uterine bleeding; IQR, interquartile range; PMB, postmenopausal bleeding. Values are n (%) unless otherwise indicated.

3.2. Expert-Rated Appropriateness Scores

Across all scenarios, the overall expert-rated appropriateness score was 3.48 ± 0.66 , with a median of 3.62 and an interquartile range of 3.00–4.00. The mean clinical appropriateness score was 3.52 ± 0.66 , the mean guideline adherence score was 3.42 ± 0.69 , the mean safety score was 3.63 ± 0.69 , and the mean completeness score was 3.33 ± 0.74 .

Overall, 21 of 60 ChatGPT responses were classified as appropriate, corresponding to 35.0%. Twenty-seven responses, corresponding to 45.0%, were classified as partially appropriate, and 12 responses, corresponding to 20.0%, were classified as inappropriate.

When analyzed according to age stratum, the mean overall score was 4.10 ± 0.41 in adolescents, 3.92 ± 0.33 in reproductive-age women, 3.56 ± 0.43 in women aged 40–44 years, 2.98 ± 0.49 in premenopausal or perimenopausal women aged 45–55 years, and 2.82 ± 0.51 in women with postmenopausal bleeding. The proportion of responses classified as appropriate was 83.3% in adolescents, 66.7% in reproductive-age women, 16.7% in women aged 40–44 years, 8.3% in women aged 45–55 years, and 0.0% in women with postmenopausal bleeding. Age-stratified expert-rated scores and appropriateness categories are presented in Table 2.

The Kruskal–Wallis test demonstrated statistically significant differences across age strata for clinical appropriateness, guideline adherence, safety, completeness, and overall score, with p values <0.001 for all domains. Dunn–Bonferroni post hoc comparisons for overall score showed significantly lower scores in the 45–55-year and postmenopausal bleeding groups than in adolescents ($p < 0.001$ for both comparisons) and reproductive-age women ($p = 0.002$ and $p < 0.001$, respectively). Other pairwise comparisons were not statistically significant after Bonferroni correction. The distribution of overall appropriateness scores across the five age strata is illustrated in Figure 1. Mean domain scores by age group are displayed as a heatmap in Figure 2.

Spearman correlation analysis showed a negative association between increasing age stratum and overall appropriateness score, with $\rho = -0.771$ and $p < 0.001$. In exploratory ordinal logistic regression, each higher age stratum was associated with lower odds of being classified in a higher appropriateness category, with an odds ratio of 0.20, 95% confidence interval 0.10–0.37, and $p < 0.001$.

Table 2A. Expert-rated performance scores by age stratum

Age stratum	N	Clinical appropriate- ness	Guideline adherence	Safety
Adolescents (13–18 y)	12	4.08 ± 0.47	4.08 ± 0.36	4.29 ± 0.45
Reproductive age (19–39 y)	12	3.96 ± 0.40	3.79 ± 0.40	4.08 ± 0.42
Late reproductive age (40–44 y)	12	3.58 ± 0.47	3.58 ± 0.47	3.67 ± 0.49

Age stratum	N	Clinical appropriateness	Guideline adherence	Safety
45–55 y pre/perimenopausal	12	3.04 ± 0.54	2.92 ± 0.56	3.12 ± 0.43
Postmenopausal bleeding	12	2.92 ± 0.51	2.75 ± 0.58	3.00 ± 0.60
Total	60	3.52 ± 0.66	3.42 ± 0.69	3.63 ± 0.69

Table 2B. Expert-rated performance scores by age stratum

Age stratum	Completeness	Overall score	Overall median [IQR]	Appropriate responses
Adolescents (13–18 y)	3.96 ± 0.54	4.10 ± 0.41	4.00 [4.00–4.19]	10/12 (83.3)
Reproductive age (19–39 y)	3.83 ± 0.44	3.92 ± 0.33	4.00 [3.72–4.06]	8/12 (66.7)
Late reproductive age (40–44 y)	3.42 ± 0.42	3.56 ± 0.43	3.75 [3.00–3.88]	2/12 (16.7)
45–55 y pre/perimenopausal	2.83 ± 0.62	2.98 ± 0.49	3.00 [2.72–3.00]	1/12 (8.3)
Postmenopausal bleeding	2.62 ± 0.61	2.82 ± 0.51	2.81 [2.50–3.03]	0/12 (0.0)
Total	3.33 ± 0.74	3.48 ± 0.66	3.62 [3.00–4.00]	21/60 (35.0)

Scores are shown as mean ± SD unless otherwise indicated. Appropriate responses were defined as overall mean score ≥4.0. Across-group Kruskal–Wallis *p* values were <0.001 for clinical appropriateness, guideline adherence, safety, completeness, and overall score. Abbreviations: IQR, interquartile range; SD, standard deviation.

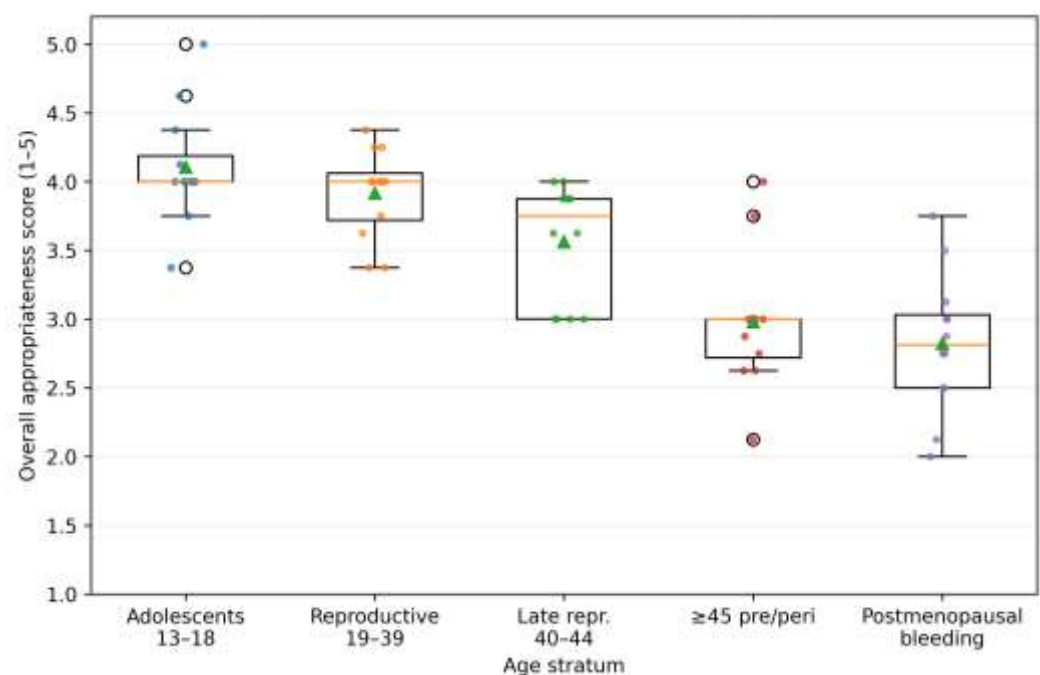


Figure 1. Distribution of overall expert-rated appropriateness scores across age groups. Boxplot showing the distribution of overall ChatGPT appropriateness scores across the five predefined age strata: adolescents aged 13–18 years, reproductive-age women aged 19–39 years, late reproductive-age women aged 40–44 years, premenopausal or perimenopausal women aged 45–55 years, and women with postmenopausal bleeding. The overall score was calculated as the mean of four expert-rated domains: clinical appropriateness, guideline adherence, patient safety, and completeness. The horizontal line within each box represents the median, the box represents the interquartile range, and whiskers indicate the range of observed values excluding outliers. Overall scores differed significantly across age groups according to the Kruskal–Wallis test (*p* < 0.001).

3.3. Endometrial Sampling Accuracy, High-Risk Recognition, and Urgency Decisions

Across reviewer-level assessments, endometrial sampling recommendations were correct in 72 of 120 evaluations, corresponding to 60.0%. Age-stratified endometrial

sampling correctness was 95.8% in adolescents, 95.8% in reproductive-age women, 62.5% in women aged 40–44 years, 20.8% in women aged 45–55 years, and 25.0% in women with postmenopausal bleeding. The difference across age strata was statistically significant, with $p < 0.001$.

Among the 20 scenarios predefined as containing high-risk or malignancy-related features, these features were correctly recognized in 19 of 40 reviewer-level assessments, corresponding to 47.5%. Recognition rates were 100.0% in adolescents, 100.0% in reproductive-age women, 25.0% in women aged 40–44 years, 12.5% in women aged 45–55 years, and 0.0% in women with postmenopausal bleeding. The difference across age strata was statistically significant, with $p < 0.001$.

Urgency decisions were correct in 116 of 120 reviewer-level assessments, corresponding to 96.7%. Age-stratified urgency decision correctness was 91.7% in adolescents, 100.0% in reproductive-age women, 95.8% in women aged 40–44 years, 95.8% in women aged 45–55 years, and 100.0% in women with postmenopausal bleeding. The difference across age strata was not statistically significant, with $p = 0.460$.

Harmful or falsely reassuring recommendations were recorded in 48 of 120 reviewer-level assessments, corresponding to 40.0%. These flags were recorded in 0.0% of adolescent assessments, 0.0% of reproductive-age assessments, 33.3% of assessments in women aged 40–44 years, 83.3% of assessments in women aged 45–55 years, and 83.3% of postmenopausal bleeding assessments. The age-stratified results for endometrial sampling accuracy, high-risk recognition, urgency decision accuracy, and harmful or falsely reassuring recommendations are presented in Table 3.

Table 3. Endometrial sampling accuracy and high-risk/malignancy recognition by age stratum

Age stratum	Endometrial sampling decision correct	High-risk/malignancy recognition correct	Urgency decision correct	Harmful/falsely reassuring flag
Adolescents (13–18 y)	23/24 (95.8)	8/8 (100.0)	22/24 (91.7)	0/24 (0.0)
Reproductive age (19–39 y)	23/24 (95.8)	8/8 (100.0)	24/24 (100.0)	0/24 (0.0)
Late reproductive age (40–44 y)	15/24 (62.5)	2/8 (25.0)	23/24 (95.8)	8/24 (33.3)
45–55 y pre/perimenopausal	5/24 (20.8)	1/8 (12.5)	23/24 (95.8)	20/24 (83.3)
Postmenopausal bleeding	6/24 (25.0)	0/8 (0.0)	24/24 (100.0)	20/24 (83.3)
Total	72/120 (60.0)	19/40 (47.5)	116/120 (96.7)	48/120 (40.0)

Values are reviewer-level assessments because each scenario was independently reviewed by two experts. High-risk/malignancy recognition was assessed only in 20 predefined high-risk scenarios, resulting in 40 reviewer-level assessments. Across-group p values were < 0.001 for endometrial sampling correctness, < 0.001 for high-risk/malignancy recognition, 0.460 for urgency correctness, and < 0.001 for harmful/falsely reassuring recommendations.

3.4. Multidimensional Performance Profile

The age-stratified performance profile across the four expert-rated domains showed mean scores of 4.08 for clinical appropriateness, 4.08 for guideline adherence, 4.29 for safety, and 3.96 for completeness in adolescents. Corresponding scores were 3.96, 3.79, 4.08, and 3.83 in reproductive-age women; 3.58, 3.58, 3.67, and 3.42 in women aged 40–44 years; 3.04, 2.92, 3.12, and 2.83 in women aged 45–55 years; and 2.92, 2.75, 3.00, and 2.62 in women with postmenopausal bleeding. The multidimensional age-stratified performance pattern is shown in Figure 3.



Figure 2. Heatmap of mean expert-rated domain scores by age group. Heatmap displaying age-stratified mean scores for the four expert-rated domains: clinical appropriateness, guideline adherence, patient safety, and completeness. Each cell represents the mean score for the corresponding domain and age group on a five-point Likert scale, where higher values indicate greater appropriateness, stronger guideline concordance, better safety, or greater completeness.

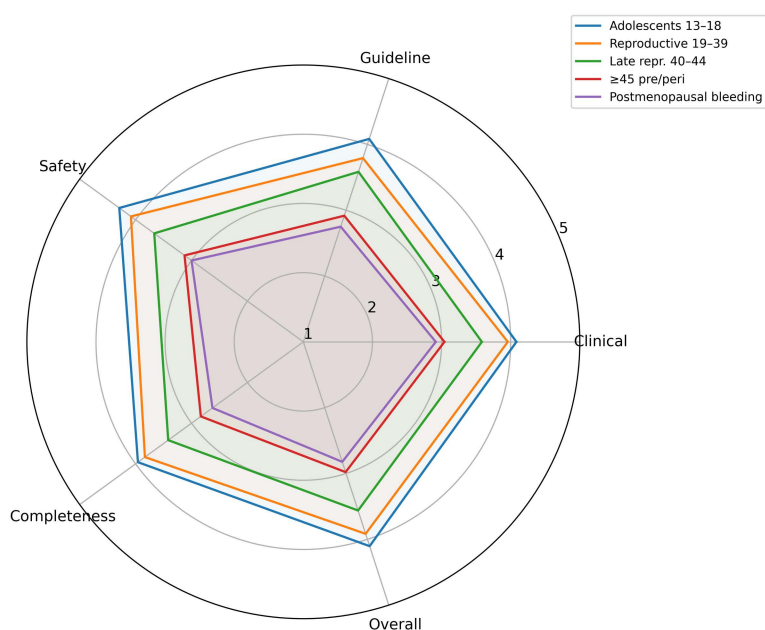


Figure 3. Multidimensional performance profile of ChatGPT responses across age groups. Radar chart showing the age-stratified performance profile of ChatGPT responses across four expert-rated domains: clinical appropriateness, guideline adherence, patient safety, and completeness. Scores are presented on a five-point scale, with higher values indicating better performance. The chart illustrates the relative pattern of domain-specific performance across the five age strata.

3.5. Inter-Rater Agreement and Reliability

Inter-rater agreement for endometrial sampling correctness was assessed using Cohen's kappa. The kappa coefficient was 0.792. Inter-rater reliability for the overall expert-rated score was assessed using the intraclass correlation coefficient. The ICC was 0.957.

4. Discussion

In this age-stratified evaluation of ChatGPT-generated recommendations for abnormal uterine bleeding, we observed a clear decline in performance with increasing patient age. Although responses were generally appropriate for adolescents and reproductive-age women, substantially lower scores were observed in women aged 45–55 years and in postmenopausal bleeding scenarios. The most clinically relevant deficiencies involved reduced accuracy of endometrial sampling recommendations, inadequate recognition of malignancy-associated features, and a higher frequency of potentially misleading or falsely reassuring advice.

The observed pattern suggests that current LLMs may perform adequately when management is primarily symptom-oriented but become less reliable when clinical decision-making requires explicit application of age-based cancer-risk thresholds.

These findings are consistent with previous studies demonstrating that LLMs can generate medically coherent and clinically plausible responses while remaining vulnerable to errors in contextual reasoning and guideline application^{8,9,17}. Although early evaluations showed that LLMs may perform well on standardized medical examinations and generate responses perceived as high quality by expert reviewers, examination performance and conversational fluency do not necessarily translate into reliable clinical triage in high-risk scenarios requiring strict guideline adherence.

Our findings also align with emerging literature in women's health and gynecologic oncology. Previous studies have shown that ChatGPT may provide useful educational information while demonstrating variable performance when recommendations require integration of multiple contextual factors¹⁰. Similarly, evaluations of LLMs against gynecologic oncology guidelines have identified inconsistent guideline concordance and incomplete recognition of oncologic risk factors^{11,12}. The present study extends these observations by demonstrating an age-dependent performance gradient in a clinical context where age directly influences malignancy risk and biopsy thresholds.

A particularly important finding was the marked reduction in endometrial sampling accuracy among women aged 45 years and older and in postmenopausal bleeding scenarios. Current guidance consistently emphasizes endometrial evaluation in these populations because of their elevated risk of hyperplasia and malignancy⁵⁻⁷. Accordingly, failure to recommend biopsy in such scenarios cannot be considered a minor omission; rather, it represents a clinically meaningful deficiency with potential implications for delayed diagnosis.

The age-dependent decline observed in malignancy-risk recognition may reflect fundamental limitations of contemporary generative AI systems. Unlike rule-based clinical decision-support systems, LLMs generate responses through probabilistic language prediction rather than explicit application of diagnostic algorithms. Consequently, recommendations may appear convincing even when essential risk factors are underweighted or overlooked. Similar concerns have been raised in gynecologic oncology and broader cancer-related applications of LLMs^{11,12}.

Another notable finding was the increasing frequency of harmful or falsely reassuring recommendations in higher-risk age groups. Although overtly dangerous advice was uncommon, omission of biopsy recommendations, underestimation of malignancy risk, or insufficient emphasis on specialist evaluation may inadvertently provide reassurance in situations requiring prompt investigation. These observations are consistent with broader concerns regarding hallucinations, missed red flags, and incomplete reasoning in medical AI systems¹³⁻¹⁶.

Interestingly, while overall safety scores remained within a moderate range, harmful or falsely reassuring recommendations were identified in 40.0% of reviewer-level assessments. This discrepancy suggests that global safety ratings may not fully capture clinically

meaningful omissions. A response may appear informative and generally appropriate while simultaneously failing to recommend a critical diagnostic intervention such as endometrial biopsy.

Beyond its clinical findings, the present study contributes methodologically to the emerging literature on AI evaluation in healthcare. Unlike many previous studies based on examination questions or educational materials, our study employed standardized age-stratified clinical scenarios independently evaluated by board-certified gynecologic oncologists. This design enabled assessment of clinically meaningful endpoints, including guideline adherence, malignancy-risk recognition, urgency classification, and endometrial sampling accuracy.

The clinical implications of these findings deserve consideration. Women experiencing abnormal bleeding increasingly use online health resources and conversational AI systems before consulting healthcare professionals. Although ChatGPT may provide reasonably appropriate information in lower-risk scenarios, our findings suggest that reliance on AI-generated recommendations may be problematic in populations at increased risk of endometrial pathology. Symptoms occurring in women aged 45 years or older or after menopause should continue to prompt formal clinical evaluation regardless of reassurance provided by conversational AI systems.

Several limitations should be acknowledged. First, the study was based on hypothetical clinical scenarios rather than real patient encounters. Although this approach allowed standardized comparisons across age groups, it cannot fully replicate the complexity of clinical practice. Second, the relatively limited number of scenarios within each age stratum may restrict generalizability. Third, only a single LLM and a single prompting strategy were evaluated. ChatGPT was selected because it remains one of the most widely used publicly accessible conversational AI systems worldwide; however, performance may differ across other models, future versions, retrieval-augmented systems, or alternative prompt-engineering strategies. In addition, each scenario was submitted to ChatGPT only once; therefore, within-model response consistency across repeated runs of the same prompt was not assessed. Finally, rapid model evolution may limit the long-term applicability of the findings.

To our knowledge, this is the first study specifically evaluating age-stratified performance of a large language model in abnormal uterine bleeding triage with particular emphasis on endometrial sampling recommendations and malignancy-risk recognition.

5. Conclusion

ChatGPT demonstrated acceptable performance in adolescent and reproductive-age AUB scenarios but showed substantially lower appropriateness, poorer endometrial sampling accuracy, and weaker malignancy-risk recognition in women aged 45–55 years and in postmenopausal bleeding scenarios. These findings suggest that current LLMs should be considered supportive informational tools rather than independent triage systems when evaluating abnormal uterine bleeding, particularly in populations at increased risk of endometrial neoplasia. Continued physician oversight remains essential when AI-generated recommendations may influence diagnostic or referral decisions.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest: The authors declare that they have no conflict of interest.

Ethics Approval: This study involved only investigator-generated hypothetical clinical scenarios and expert evaluation of artificial intelligence-generated responses and did not include human participants, real patient data, biological materials, clinical records, images, institutional health data, or identifiable personal information. The study protocol was approved by the İzmir City Hospital Ethics Committee on June 17, 2026 (decision no: 2026/459).

Informed Consent: Not applicable, because the study did not involve human participants, identifiable patient information, or real patient data.

Clinical Trial Registration: Not applicable.

Data Availability: The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

Author Contributions (CRediT): Conceptualization: S.Ö., M.S.; Methodology: S.Ö., M.S.; Investigation: S.Ö., E.G.Ö.; Data curation: S.Ö., E.G.Ö.; Formal analysis: S.Ö.; Validation: E.G.Ö., M.S.; Writing - original draft: S.Ö.; Writing - review and editing: S.Ö., E.G.Ö., M.S.; Supervision: M.S. All authors read and approved the final version of the manuscript.

References

1. Marnach ML, Laughlin-Tommaso SK. Evaluation and management of abnormal uterine bleeding. *Mayo Clin Proc.* 2019;94(2):326-335. doi:10.1016/j.mayocp.2018.12.012. PMID: 30711128.
2. Munro MG, Critchley HOD, Broder MS, Fraser IS. FIGO classification system (PALM-COEIN) for causes of abnormal uterine bleeding in nongravid women of reproductive age. *Int J Gynaecol Obstet.* 2011;113(1):3-13. doi:10.1016/j.ijgo.2010.11.011. PMID: 21345435.
3. Munro MG, Critchley HOD, Fraser IS. The two FIGO systems for normal and abnormal uterine bleeding symptoms and classification of causes of abnormal uterine bleeding in the reproductive years: 2018 revisions. *Int J Gynaecol Obstet.* 2018;143(3):393-408. doi:10.1002/ijgo.12666. PMID: 30198563.
4. Jain V, Munro MG, Critchley HOD. Contemporary evaluation of women and girls with abnormal uterine bleeding: FIGO Systems 1 and 2. *Int J Gynaecol Obstet.* 2023;162 Suppl 2(Suppl 2):29-42. doi:10.1002/ijgo.14946. PMID: 37538019.
5. Papakostantinou E, Adonakis G. Management of pre-, peri-, and post-menopausal abnormal uterine bleeding: when to perform endometrial sampling? *Int J Gynaecol Obstet.* 2022;158(2):252-259. doi:10.1002/ijgo.13988. PMID: 34669187.
6. American College of Obstetricians and Gynecologists. ACOG Committee Opinion No. 734: The role of transvaginal ultrasonography in evaluating the endometrium of women with postmenopausal bleeding. *Obstet Gynecol.* 2018;131(5):e124-e129. doi:10.1097/AOG.0000000000002631. PMID: 29683909.
7. Clarke MA, Long BJ, Del Mar Morillo A, Arbyn M, Bakkum-Gamez JN, Wentzensen N. Association of endometrial cancer risk with postmenopausal bleeding in women: a systematic review and meta-analysis. *JAMA Intern Med.* 2018;178(9):1210-1222. doi:10.1001/jamainternmed.2018.2820. PMID: 30083701.
8. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health.* 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198. PMID: 36812645.
9. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ.* 2023;9:e45312. doi:10.2196/45312. PMID: 36753318.
10. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med.* 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838. PMID: 37115527.
11. Recker F, Neubauer R, Wittek AF, Scholten N. Large language models and women's health: a digital companion for informed decision-making. *Arch Gynecol Obstet.* 2025;312(3):663-670. doi:10.1007/s00404-025-08065-9. PMID: 40542853.
12. Reicher L, Lutsker G, Michaan N, Grisaru D, Laskov I. Exploring the role of artificial intelligence, large language models: comparing patient-focused information and clinical decision support capabilities to the gynecologic oncology guidelines. *Int J Gynaecol Obstet.* 2025;168(2):419-427. doi:10.1002/ijgo.15869. PMID: 39161265.
13. Gungor ND, Esen FS, Tasci T, Gungor K, Cil K. Navigating gynecological oncology with different versions of ChatGPT: a transformative breakthrough or the next black box challenge? *Oncol Res Treat.* 2025;48(3):102-111. doi:10.1159/000543173. PMID: 39689699.
14. Rosati A, Criscione M, Lilli L, Petrillo M, Fagotti A, et al. Natural language processing as consultation service platform or clinical decision support system in gynecologic oncology: a systematic review. *Int J Gynecol Cancer.* 2026;36(4):104558. doi:10.1016/j.ijgc.2026.104558. PMID: 41819636.
15. Freyer O, Wiest IC, Kather JN, Gilbert S. A future role for health applications of large language models depends on regulators enforcing safety standards. *Lancet Digit Health.* 2024;6(9):e662-e672. doi:10.1016/S2589-7500(24)00124-9. PMID: 39179311.
16. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel).* 2023;11(6):887. doi:10.3390/healthcare11060887. PMID: 36981544.
17. Lima HA, Trocoli-Couto PHFS, Moazzam Z, et al. Quality assessment of large language models' output in maternal health. *Sci Rep.* 2025;15(1):22474. doi:10.1038/s41598-025-03501-x. PMID: 40593918.