Assessing the Knowledge of ChatGPT in Answering Questions Regarding Female Urology

Hakan Cakir[1]*, Ufuk Caglar[2], Ahmet Halis[2], Omer Sarilar[2], Huseyin Burak Yazili[2], Faruk Ozgor[2]

Purpose: With the recent increase in the use of artificial intelligence in the medical field, this study aimed to evaluate the accuracy and adequacy of ChatGPT's responses to questions related to female urology.

Methods: Intensive internet research was performed to prepare a frequently asked question (FAQs) list. Scientific questions were created in accordance with the European Urology Association (EAU) Non-neurogenic Female Lower Urinary Tract Symptoms Guidelines, EAU Chronic Pelvis Pain Guidelines, and EAU Neuro-Urology Guidelines. All answers by ChatGPT were analysed by two experienced urologists and each answer was scored between 1 and 4 by the physicians. A score of 1 was the highest and showed that the answer was completely true and sufficient. The reproducibility of ChatGPT answers was evaluated by asking each question twice using two different computers.

Results: A total of 96 (97.0%) ChatGPT answers about female urology were accurate and sufficient, and categorized as grade 1. Additionally, two (2.0%) answers were scored as grade 2, and one answer (1.0%) was scored as grade 3. None of ChatGPT's responses about female urology were classified as grade 4. In total, 83 questions were prepared according to EAU guidelines recommendations, and ChatGPT gave complete accurate and satisfactory answers for 68 (82.9%) questions. The reproducibility rate was highest for ChatGPT answers for questions related to urinary incontinence, pelvic organ prolapses, and pelvic pain syndromes, and reproducibility rate was 100% for each subgroup. The reproducibility rate for ChatGPT answers was lowest for CPG questions (84.1%).

Conclusion: For the first time our study revealed that ChatGPT had an excellent accuracy rate in answering questions related to female urology with 97% success rate. In addition, the outcomes of this study showed that ChatGPT accurately and satisfactorily answered 82.9% of questions about female urology based on EAU guidelines.

Keywords: artificial intelligence; ChatGPT; guideline; female urology

INTRODUCTION

Contrary to popular belief, urological diseases are common in women, and urinary incontinence affects 50% of adult women, including up to 80% of women over 70. Pelvic organ prolapse is detected in 2 out of every 5 women[1]. Moreover, current studies show that the prevalence of interstitial cystitis/bladder pain syndrome (IC/BPS), pelvic pain syndrome and female genital fistulas has been increasing in recent years[2,3]. Although the diseases mentioned above are not life-threatening in most cases, negative effects of these disorders on the patient's work life and social life and on the healthcare system are well-known[4]. Increasing public awareness of any disease will enable early diagnosis and also increase patient compliance with treatment and follow-up processes. At the present time, many people use internet-based applications to gain knowledge about health conditions and diseases[5].

ChatGPT is an artificial intelligence multiple language based chatbot which provides humanlike conversations (OpenAI, California, USA)[6]. Nowadays, ChatGPT has become an important source of information frequently used by patients, and the use of ChatGPT in the field of healthcare stands out as one of the most important discussion topics. Caglar and colleagues analyzed the knowledge of ChatGPT about pediatric urology, and found that ChatGPT answered all questions with 92% complete accuracy rate[7]. Rahimli et al. conducted a study evaluating the accuracy of answers provided by artificial intelligence applications to questions about pelvic organ prolapse. For the responses generated by ChatGPT, they found an accuracy rate of 93.3%, a completeness rate of 66.7%, and a precision rate of 66.7% [8]. In addition, Choueka et al. showed that ChatGPT can help generate ideas for scientific studies on urogynecology[9].

Until today, a limited number of studies have analysed the knowledge of ChatGPT about different medical conditions. To our knowledge, no study has investigated the accuracy of ChatGPT answers about female urology. In the present study, for the first time the knowledge of ChatGPT about female urology was analysed.

[1]Jingzhou Hospital Affiliated to Yangtze University, Jingzhou, 434020, China
[2]The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, 230001, China
[3]The People's Hospital of Longhua, Shenzhen, 518109, China
YW and ZZ contribute equally to this study.
*Correspondence: Department of Urology, The People's Hospital of Longhua, 38 Jinglongjianshe Road, Longhua District, Shenzhen, Guangdong 518109, China. Tel: +86 755 27741585. Fax: +86 755 29407559. E-mail: dengqiong1987@smu.edu.cn
Received December 2023 & Accepted June 2024

Table 1. Grade of responses by ChatGPT to questions related to female urology

	Grade 1	Grade 2	Grade 3	Grade 4
Frequently asked questions (n=99)	96 (97.0%)	2 (2.0%)	1 (1.0%)	-
Urinary incontinence (n=19)	18 (94.7%)	1 (5.3%)	-	-
Pelvic organ prolapse (n=20)	20 (100.0%)	-	-	-
IC/BPS (n=20)	19 (95.0%)	1 (5.0%)	-	-
Pelvic pain syndromes (n=20)	20 (100.0%)	-	-	-
Female genital fistulas (n=20)	19 (95.0%)	-	1 (5.0%)	-
EAU guideline recommendations (n=82)	68 (82.9%)	8 (9.8%)	6 (7.3%)	-

Grade 1: Completely correct, Grade 2: Correct but insufficient, Grade 3: Misleading information as well as correct information, Grade 4: Completely incorrect, EAU: European Association of Urology, IC/BPS: Interstitial cystitis/bladder pain syndrome

MATERIALS AND METHODS

In the present study, intensive internet research was performed using official websites of hospitals, health institutions, healthcare organizations, and personal healthcare providers to prepare a frequently asked question (FAQs) list. Additionally, patients' comments and frequent inquiries on popular social media platforms including Twitter, Facebook, YouTube, Instagram etc. were analysed. For this analysis, a keyword search strategy was employed to identify relevant posts and comments. Keywords related to female urology, such as "urinary incontinence," "pelvic organ prolapse," "IC/BPS," "pelvic pain syndrome," and "female genital fistulas" were used. Posts and comments were then manually reviewed to extract frequently asked questions and common concerns. All questions prepared from the aforementioned sources are documented in Supplementary File 1. In addition, scientific questions were created in accordance with the European Urology Association (EAU) Non-neurogenic Female Lower Urinary Tract Symptoms Guidelines, EAU Chronic Pelvis Pain Guidelines, and EAU Neuro-Urology Guidelines. All questions based on these three guidelines are listed in Supplementary File 2 as clinical practice guideline

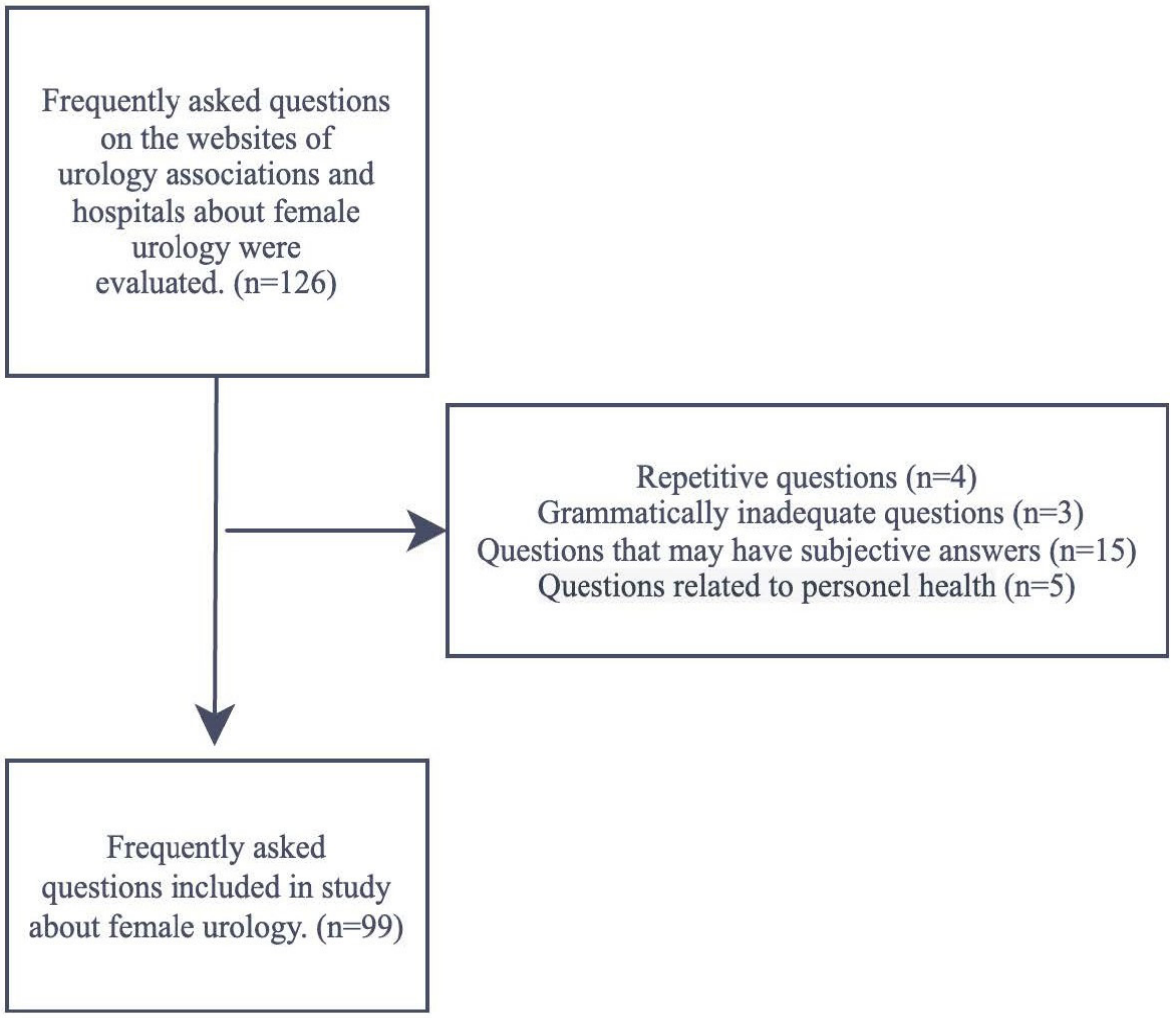

Figure 1. Flowchart for questions included in the study

Table 2. Performance measures of the information generated by ChatGPT: F1 score, precision score, recall score.

	Precision	Recall	F1 score
Frequently asked questions (n=99)			
Urinary incontinence (n=19)	0.95	1.0	0.97
Pelvic organ prolapse (n=20)	1.0	1.0	1.0
IC/BPS (n=20)	0.95	1.0	0.97
Pelvic pain syndromes (n=20)	1.0	1.0	1.0
Female genital fistulas (n=20)	1.0	0.95	0.97
EAU guideline recommendations (n=82)	0.89	0.92	0.91

EAU: European Association of Urology, IC/BPS: Interstitial cystitis/bladder pain syndrome

(CPG) questions. Also, ChatGPT answers to CPG questions were evaluated in accordance with the aforementioned guidelines.

Advertising questions, similar questions, questions containing obvious grammatical errors, questions requiring subjective responses, and personal health questions were excluded from the study. In total, 99 FAQs were prepared, comprising 19 questions about urinary incontinence, 20 questions about pelvic organ prolapse, 20 questions about IC/BPS, 20 questions about pelvic pain syndrome, and 20 questions about female genital fistulas. Moreover, a total of 82 questions were created in accordance with EAU guidelines.

For the present study, the ChatGPT free version was used, and all questions were asked on 1st September 2023. The version of ChatGPT used was GPT-3.5. To ensure consistency and minimize bias in responses, a standardized prompt format was employed for all questions. Each question was carefully crafted to be clear and specific, avoiding ambiguity and subjective language. The prompts were designed to reflect typical patient inquiries and clinical questions, aligned with the structure and terminology used in the European Urology Association (EAU) guidelines. Additionally, all questions were asked in natural language without leading or suggestive wording to ensure that ChatGPT's responses were as accurate and unbiased as possible. All answers by ChatGPT were analysed by two experienced urologists with a minimum of 10 years of experience, and each answer was scored between 1 and 4 by the physicians. A score of 1 was the highest and showed that the answer was completely true and sufficient. If the ChatGPT answer was accurate but insufficient, the answer was scored as 2. Answers with combinations of accurate and misleading information were categorized a score of 3, and completely incorrect answers were given a score of 4. If the two urologists evaluated the same answer in different categories, the score given to the answer was re-evaluated and determined by the common decision of the urologists.

The reproducibility of ChatGPT answers was evaluated by asking each question twice using two different computers. The same score for the same question on different computers was accepted as positive for ChatGPT reproducibility. Ethics committee approval was not required as patient data were not used.

Statistical Analysis

Excel Version 16 (Microsoft Corporation, USA) was used for statistical analysis. The questions were analysed separately as FAQs and CPG questions. The scores for ChatGPT answers are given as percentages. Two ChatGPT answers with different scores for the same question were accepted as negative in terms of reproducibility. Intra-class correlation was based on a two-way random effects model with type consistency. The correlation was classified as poor (ICC < 0.40), fair to good (ICC = 0.40 - 0.75), and excellent (ICC > 0.75). Statistical measures such as precision, recall and F1

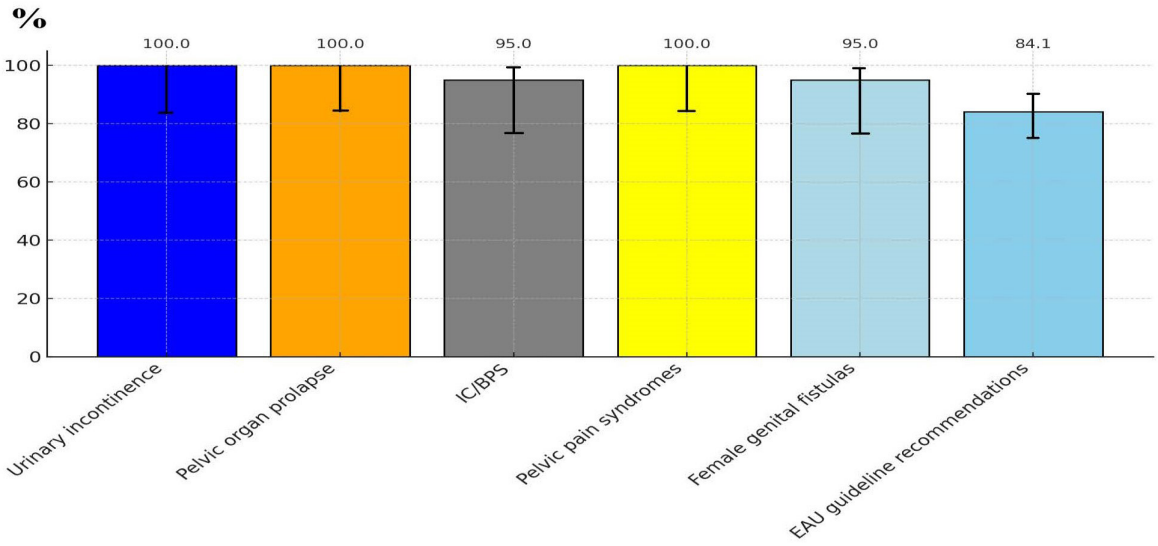

Figure 2. Similarity rates of answers to questions

score were used to evaluate the performance of the models. The responses were evaluated against the reference responses and the following parameters were calculated: True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). Formulas for performance evaluation:
Precision = TP / (TP + FP),
Recall = TP / (TP+FN),
F1 score = 2 x (precision x recall) / (precision + recall).

RESULTS

In the end, 126 FAQs about female urology were evaluated in accordance with the study design. Four repetitive questions, 3 grammatically inadequate questions, 15 questions without objective answers, and 5 questions related to personal health were excluded from the study. The flowchart of FAQs enrolled in the study is presented in **Figure 1**.

A total of 96 (97.0%) ChatGPT answers about female urology were accurate and sufficient, and categorized as grade 1. Additionally, two (2.0%) answers were scored as grade 2, and one answer (1.0%) was scored as grade 3. None of ChatGPT's responses about female urology were classified as grade 4. The complete accuracy rate of ChatGPT answers was 94.7% (18 out of 19 questions), 100% (20 out of 20 questions), 95% (19 out of 20 questions), 100% (20 out of 20 questions) and 95% (19 out of 20 questions) for the topics of urinary incontinence, pelvic organ prolapses, IC/BPS, pelvic pain syndrome, and female genital fistulas, respectively. The answers that were not completely correct were related to the diagnosis of urinary incontinence, the diagnosis of IC/PBS, and the treatment of genital fistula.

In total, 83 questions were prepared according to EAU guidelines recommendations, and ChatGPT gave complete accurate and satisfactory answers for 68 (82.9%) questions. Eight (9.8%) ChatGPT answers to CPG questions were scored as grade 2, and six (7.3%) ChatGPT answers to CPG questions were classified as grade 3. The grades for answers by ChatGPT to questions related to female urology are summarized in Table 1. In the analysis of the intraclass correlation coefficient for the measurements conducted by the observers, the consistency was found to be excellent (ICC: 0.925; %95 CI: 0.877 – 0.951; *p*: 0.001).

Performance metrics (F1 score, precision score, recall score) of information generated by ChatGPT are shown in Table 2. In the frequently asked questions, high success was determined with the F1 score between 0.97 and 1 in different disease categories. F1 score for EAU guideline questions was calculated as 0.91.

Similarity rates of ChatGPT answers to questions about female urology are shown in **Figure 2**. The reproducibility rate was highest for ChatGPT answers for questions related with urinary incontinence, pelvic organ prolapses, and pelvic pain syndromes, and reproducibility rate was 100% for each subgroup (%95 CI: 83.2 – 100, %95 CI: 86.8 – 100, and % 95 CI: 86.8 – 100; respectively). The reproducibility rate for ChatGPT answers was lowest for CPG questions (84.1%, %95 CI: 76.3 – 89.9).

DISCUSSION

While artificial intelligence (AI) has begun to be used more in all areas of life, the pros and cons of AI are one of the most interesting topics. With the use of AI, screening tests can be used more effectively, early diagnosis of patients and patient compliance with treatment processes can be increased. Also, AI could prevent unnecessary admissions to hospitals and reduce the burden on the healthcare system[10]. However, the competence and reliability of AI are still under debate in medicine, and studies about the use of AI in the field of healthcare continue. Thus, this research was planned to evaluate the knowledge of ChatGPT about female urology. Our results showed that ChatGPT had an excellent accuracy rate in answering questions related to female urology with 97% success rate. Moreover, the findings of the present study demonstrate that ChatGPT accurately and satisfactorily answered 82.9% of questions based on EAU guidelines about female urology. Lastly, ChatGPT achieved over 95% reproducibility for FAQs and 84.1% reproducibility for CPG questions.

A significant portion of the content uploaded to the internet in the field of health is not evaluated for accuracy and reliability. Yuksel and Cakmak conducted a study to determine the quality of YouTube videos about COVID-19 and pregnancy. They found that despite the high view rate of videos, most videos had inaccurate and insufficient content[11]. In another study, Alsyouf and colleagues investigated the quality and reliability of content uploaded to social media platforms about genitourinary cancers, and the authors stated that there was almost 30 times higher misleading information than accurate data[12]. In contrast to the aforementioned data, Bulck and Moons investigated the performance of ChatGPT in answering questions related to cardiac disease. The authors emphasized that ChatGPT gave complete accurate responses for 85% of inquiries[13]. This study evaluated the knowledge of ChatGPT in answering questions about female urology for the first time, and the outcomes of the present study demonstrate that ChatGPT had an excellent success rate in answering questions about female urology. Many social media applications do not have strict checking mechanisms for uploaded content in terms of accuracy and reliability. Unlike these platforms, ChatGPT can access various internet-based sources including newspapers, scientific papers and books. We believe that the ability of ChatGPT to screen numerous internet-based sources is associated with the high accuracy and reliability rate of ChatGPT answers.

Guidelines in urology include important and specific suggestions for urologists, and recommendations were created in accordance with numerous scientific papers including original studies, reviews, and meta-analysis. ChatGPT answers to questions involving intensive scientific knowledge can be sophisticated. In a study by Caglar et al., which investigated the knowledge of ChatGPT in answering questions about paediatric urology guidelines, the accuracy rate for ChatGPT was 93.6% for questions prepared according to EAU paediatric urology guidelines[7]. In another study, Antaki and colleagues analysed the performance of ChatGPT in ophthalmology residence exams, and ChatGPT received 55.8% on the exam. Although the ChatGPT's score was below the percentage of correct answers to normal questions, it is similar to the average score for ophthalmology residents[14]. In the present study, ChatGPT gave totally correct answers for 4 out of every 5 questions prepared according to EAU female urology guidelines.

This study analysed a certain time interval, which could be accepted as a limitation of the study. New knowledge about female urology is continuously uploaded and updated on the internet. Secondly, the performance of ChatGPT in answering inquiries about female urology was analysed only in the English language. However, English is the most common language in academic areas and on websites. The quality and reliability of ChatGPT answers in less common languages may be evaluated by further studies. In this study, the accuracy of ChatGPT answers was evaluated from the physician's perspective. The understandability of the answers given by ChatGPT for the public may be the subject of a different study.

CONCLUSIONS

For the first time, our study revealed that ChatGPT had an excellent accuracy rate in answering questions related to female urology with 97% success rate. In addition, the outcomes of this study showed that ChatGPT accurately and satisfactorily answered 82.9% of questions about female urology based on EAU guidelines.

CONFLICT OF INTEREST

The authors report no conflict of interest.

REFERENCES

1. Anger JT, Saigal CS, Pace J, Rodríguez LV, Litwin MS, Urologic Diseases of America Project. True prevalence of urinary incontinence among female nursing home residents. Urol. 2006;67:281-7.
2. Chen A, Shahiyan RH, Anger JT. Interstitial cystitis/bladder pain syndrome treatment: a systematic review of sexual health outcomes. Sex Med Rev. 2022;10:71-6.
3. Tasnim N, Bangash K, Amin O, Luqman S, Hina H. Rising trends in iatrogenic urogenital fistula: A new challenge. Int J Gynaecol Obstet. 2020;148:33-6.
4. Abufaraj M, Xu T, Cao C, Siyam A, Isleem U, Massad A, Soria F, Shariat SF, Sutcliffe S, Yang L. Prevalence and trends in urinary incontinence among women in the United States, 2005–2018. Am J Obstet Gynecol. 2021;225:166-e1.
5. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res. 2021;23:e17917..
6. Samaan JS, Yeo YH, Rajeev N, et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. Obes Surg. 2023 Apr 27:1-7.
7. Caglar U, Yildiz O, Meric A, et al. Evaluating the performance of ChatGPT in answering questions related to pediatric urology. J Pediatr Urol. 2023:S1477-5131(23)00318-2.
8. Rahimli Ocakoglu S, Coskun B. The Emerging Role of AI in Patient Education: A Comparative Analysis of LLM Accuracy for Pelvic Organ Prolapse. Med Princ Pract. Published online March 25, 2024. doi:10.1159/000538538
9. Choueka D, Tabakin AL, Shalom DF. ChatGPT in Urogynecology Research: Novel or Not?. Urogynecol. Published online March 25, 2024. doi:10.1097/SPV.0000000000001505
10. Zhou Z. Evaluation of ChatGPT's Capabilities in Medical Report Generation. Cureus. 2023;15:e37589.
11. Yuksel B, Cakmak K. Healthcare information on YouTube: Pregnancy and COVID-19. Int J Gynaecol Obstet. 2020;150:189-93.
12. Alsyouf M, Stokes P, Hur D, Amasyali A, Ruckle H, Hu B. 'Fake News' in urology: evaluating the accuracy of articles shared on social media in genitourinary malignancies. BJU Int. 2019;124:701-6.
13. Van Bulck L, Moons P. Response to the Letter to the Editor on: Dr. ChatGPT in Cardiovascular Nursing: A Deeper Dive into Trustworthiness, Value, and Potential Risks . Eur J Cardiovasc Nurs. 2023;zvad049.
14. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. Ophthalmol Sci. 2023;3:100324.