Received: 30 April 2024 | Revised: 30 May 2024 | Accepted: 14 June 2024 | Published online: 30 June 2024

DOI: 10.1002/ijgo.15751

CLINICAL ARTICLE

Gynecology

# The future of patient education: A study on AI-driven responses to urinary incontinence inquiries

Reut Rotem[1,2] | Omri Zamstein[3] | Misgav Rottenstreich[2] | Orfhlaith E. O'Sullivan[1] | Barry A. O'reilly[1] | Adi Y. Weintraub[3]

[1]Department of Urogynaecology, Cork University Maternity Hospital, Cork, Ireland

[2]Department of Obstetrics and Gynecology, Shaare Zedek Medical Center, Affiliated with the Hebrew University School of Medicine, Jerusalem, Israel

[3]Department of Obstetrics and Gynecology, Soroka University Medical Center, Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel

**Correspondence**
Misgav Rottenstriech, Department of Obstetrics and Gynecology, Shaare Zedek Medical Center, 12 Bayit Street, Jerusalem 91031, Israel.
Email: misgavr@gmail.com

## Abstract

**Objective:** To evaluate the effectiveness of ChatGPT in providing insights into common urinary incontinence concerns within urogynecology. By analyzing the model's responses against established benchmarks of accuracy, completeness, and safety, the study aimed to quantify its usefulness for informing patients and aiding healthcare providers.

**Methods:** An expert-driven questionnaire was developed, inviting urogynecologists worldwide to assess ChatGPT's answers to 10 carefully selected questions on urinary incontinence (UI). These assessments focused on the accuracy of the responses, their comprehensiveness, and whether they raised any safety issues. Subsequent statistical analyses determined the average consensus among experts and identified the proportion of responses receiving favorable evaluations (a score of 4 or higher).

**Results:** Of 50 urogynecologists that were approached worldwide, 37 responded, offering insights into ChatGPT's responses on UI. The overall feedback averaged a score of 4.0, indicating a positive acceptance. Accuracy scores averaged 3.9 with 71% rated favorably, whereas comprehensiveness scored slightly higher at 4 with 74% favorable ratings. Safety assessments also averaged 4 with 74% favorable responses.

**Conclusion:** This investigation underlines ChatGPT's favorable performance across the evaluated domains of accuracy, comprehensiveness, and safety within the context of UI queries. However, despite this broadly positive reception, the study also signals a clear avenue for improvement, particularly in the precision of the provided information. Refining ChatGPT's accuracy and ensuring the delivery of more pinpointed responses are essential steps forward, aiming to bolster its utility as a comprehensive educational resource for patients and a supportive tool for healthcare practitioners.

KEYWORDS

artificial intelligence-generated responses, ChatGPT, expert opinions, quality evaluation, urinary incontinence, Urogynecology

## 1 | INTRODUCTION

The prevalence of urinary incontinence (UI) significantly affects around half of the global adult female population, with a noticeable increase in incidence among the elderly demographic.[1] UI is categorized into various types, with the three most prevalent being stress, urgency, and overflow incontinence. Individuals who exhibit symptoms from more than one category are diagnosed with mixed UI.[2] Although UI does not directly contribute to an increase in mortality rates, it profoundly impacts the quality of life of women. The range of adverse effects stemming from this condition includes, but is not limited to, depressive symptoms, diminished work productivity, sexual dysfunction, decreased mobility, and an elevated risk of associated health complications such as infections and a higher propensity for falls.[3–5]

This significant impact on individuals' well-being is further exacerbated by a widespread reluctance among those affected to seek professional consultation. Often, this reluctance is rooted in embarrassment, a lack of knowledge concerning available treatments, and an apprehension towards medical interventions.[6]

The evolution of technology, particularly the internet and artificial intelligence (AI), has revolutionized access to medical knowledge, enabling patients to interact more easily with healthcare professionals, explore available treatment modalities, and share their experiences. Among these technologic advancements, AI-powered language models have emerged as a potential tool to assist patients by providing explanations, responding to inquiries, and delivering tailored information based on individual inputs.[7] Nonetheless, these models also present limitations, including a potential lack of comprehensive understanding of complex medical scenarios, inherent biases, and the risk that an overreliance on digital resources could deter individuals from seeking necessary in-person medical consultations.[8]

The effectiveness of AI-powered language models in diagnostics and management across various medical disciplines has been subject to scrutiny, revealing both semantic benefits and concerns regarding the medical adequacy of the responses provided.[9,10] Notably, Grünebaum et al.[11] have underscored the utility of ChatGPT, an AI chatbot made available by OpenAI in November 2022,[12] in offering preliminary information on a broad spectrum of topics within obstetrics and gynecology. Moreover, subsequent research into ChatGPT's performance in addressing inquiries related to endometriosis has yielded satisfactory results.[13] However, the specific issue of female UI in the context of the accuracy of AI-driven language models remains largely unexplored. This research gap highlights the importance of evaluating the reliability of the information provided by these models regarding UI, thereby assessing their potential to enhance patient understanding and management of this widespread condition.

## 2 | MATERIALS AND METHODS

In order to examine the precision of ChatGPT's responses within the specialized domain of UI, an 11-question data set was crafted. These questions were carefully selected for their relevance to the field and their potential impact on clinical outcomes. The questionnaire was developed by a fellow in Urogynecology (RR) and an experienced specialist (AYW), while exercising their professional expertise to ensure that the queries accurately reflected the concerns and information needs of individuals seeking urogynecologic advice.

For generating answers to the selected questions, we used the most current iteration of the ChatGPT language model (Table S1). A third-party investigator, not previously involved in the question development process, was tasked with inputting the queries into the ChatGPT platform. This approach was adopted to guarantee impartiality. The investigator was given explicit instructions to solicit the AI for precise and understandable responses, with the directive: "I am seeking advice on urogynecologic issues. Please provide specific and clear answers."

The validation of the AI-generated answers was carried out by a panel of 37 urogynecologists from diverse geographical regions, including Asia, the Middle East, Europe, North America, and South America. These experts, each with extensive experience in both hospital and community healthcare settings, evaluated the responses according to three key criteria: (1) accuracy, (2) comprehensiveness, and (3) safety. A five-point Likert scale was used for the evaluation, with 1 indicating Strongly disagree, 2 indicating Disagree, 3 indicating Neutral, 4 indicating Agree, and 5 indicating Strongly agree. The complete questionnaire is detailed in Appendix S1. Responders were blinded to the fact that the answers were provided by ChatGPT. To assess the credibility of our participants, we intentionally modified one of ChatGPT's responses to contain an incorrect answer, serving as a control to gauge whether participants could accurately identify the discrepancy.

The analysis stage focused on processing the feedback received from the urogynecologists. This involved computing average scores to depict the general consensus on the responses. In addition, the percentage of specialists providing a "favorable evaluation" (ratings of 4 or 5) for each criterion in the responses was calculated. The score of 4 was chosen as favorable based on a previous paper that assessed a similar topic in obstetrics.[14] This measure was adopted to examine the extent of agreement on the responses' adherence to the principles of accuracy, comprehensiveness, and safety.

The present study was conducted with due regard for ethical standards, precluding the need for approval from an ethics committee because of the non-involvement of patient data or personal data. Measures were put in place to ensure that all participant data remained anonymous, safeguarding individual privacy and mitigating the risk of disseminating misleading or harmful health information. Importantly, no study participant had any affiliations with the developers of ChatGPT, ensuring an unbiased evaluation process.

## 3 | RESULTS

In the present study, 37 urogynecologists, out of an invited pool of 50, contributed their expertise, achieving a notable 74% response rate. The survey revealed a balanced gender distribution among the urogynecologists, with 52.8% male and 47.2% female professionals.

A significant majority (75%), had furthered their specialization through designated fellowship programs. Except for four respondents (10.8%), all participants worked at least part-time in the public sector. Notably, most responders ($n = 20$, 54%) had more than 10 years of experience. The responders were from geographically diverse locations, specifically, from the Middle East ($n = 10$), Europe ($n = 8$), South America ($n = 7$), North America ($n = 5$), and Asia ($n = 4$).

Table 1 provides a comprehensive overview of ChatGPT's performance across various criteria. Initially, we aggregated the ratings for all aspects, achieving an average score of 4.0. Furthermore, 73% of all ratings provided were 4 or higher, after excluding the "control" question.

Subsequent in-depth analysis of ChatGPT's cumulative responses involved calculating mean scores for each specific criterion. This evaluation showed an average Likert scale rating of 3.9 (with 71% scoring 4 or above) for accuracy, 4.0 (and 74% scoring 4 or above) for comprehensiveness, and 4.0 (with 74% scoring 4 or above) for safety.

Table 2 describes the variance in ratings for different ChatGPT responses. Our control question, notably, had an average rating of 2.6 with only 6% of responders awarding a score of 4 or higher. The response concerning daily management of overactive bladder/urge UI was rated highest (4.3), and the question about risks associated with UI during pregnancy received the lowest score (3.7).

Figure 1 illustrates the distribution of ratings based on criteria across ChatGPT's answers. Except for our control question, every query scored above 3 on average. ChatGPT did well particularly in comprehensiveness and safety, where 74% of responses received a score of 4 or higher. Nevertheless, there was a slight shortfall in accuracy, with only 71% of the responses scoring 4 or above.

**TABLE 1** Comprehensive evaluation of ChatGPT's performance.

| Performance | Average score | Rate of respondents giving a rating of ≥4 (%) |
|---|---|---|
| All questions | 4.0 | 73 |
| All questions regarding accuracy | 3.9 | 71 |
| All questions regarding comprehensiveness | 4.0 | 74 |
| All questions regarding safety | 4.0 | 74 |

**TABLE 2** Detailed assessment of ChatGPT responses across criteria.

| Response | Average score | Rate of respondents giving a rating of 4 (%) |
|---|---|---|
| SUI causes | 4.0 | 75 |
| SUI non-surgical treatments | 4.0 | 71 |
| SUI surgical success | 3.9 | 69 |
| SUI recurrence prevention | 4.1 | 77 |
| UUI/OAB daily management | 4.3 | 82 |
| OAB progression risk | 4.1 | 77 |
| OAB/UUI surgical options | 4.0 | 73 |
| SUI/UUI/OAB exercise guidelines | 3.8 | 63 |
| Pregnancy, SUI/UUI/OAB risks | 3.7 | 61 |
| SUI/UUI/OAB sexual health | 4.2 | 83 |
| SUI Pre-surgery assessment (control question) | 2.6 | 6 |

Abbreviations: OAB, overactive bladder; SUI, stress urinary incontinence; UUI, urge urinary incontinence.

## 4 | DISCUSSION

In exploring ChatGPT's capabilities in urogynecology, our study evaluated insights from 37 urogynecologists worldwide, achieving a 74% response rate. The findings demonstrated consistent accuracy, comprehensiveness, and safety, underscoring ChatGPT's potential as a supportive tool for patient education and healthcare provider assistance.

A key methodologic element involved introducing a control question, subtly modified to test our participants' evaluative accuracy. The noticeably lower ratings this control question received reinforced the validity of our assessment process, affirming the judgment of our specialist responders. Moreover, to address the dynamic nature of AI models, which may produce varying responses over time, we ensured that all evaluations were based on responses generated at a single time point, though it is acknowledged that different phrasings could yield different results. This anonymity was crucial in mitigating potential bias, guaranteeing that evaluations were objective and solely focused on content quality. Such an approach strengthened the credibility of our findings, affirming that the positive acceptance of ChatGPT's responses was based on their merit rather than on preconceived notions about AI's capabilities.

Although the answers generally received positive approval with an average score of 4.0 and mostly favorable ratings, a detailed analysis revealed some complexities. ChatGPT demonstrated proficient management of UI inquiries, consistent with previous findings in obstetrics and gynecology.[12–14] However, a nuanced examination of accuracy, comprehensiveness, and safety identified essential areas for refinement, particularly concerning minor discrepancies in accuracy.

Unlike many other conditions in obstetrics and gynecology, UI carries not only a clinical burden but a significant emotional one as well. Many women report feelings of embarrassment,[15] that may often prevent them from seeking the professional help they need.[16,17] This highlights the unique value of AI platforms such as ChatGPT, providing a non-judgmental, accessible platform for women to seek initial information and guidance, serving as a crucial first step for those hesitant to approach healthcare providers about such sensitive issues.

At the core of our investigation is a big question that goes beyond the usefulness of AI in health care. It raises the question of whether the increasing use of AI tools like ChatGPT decreases the importance

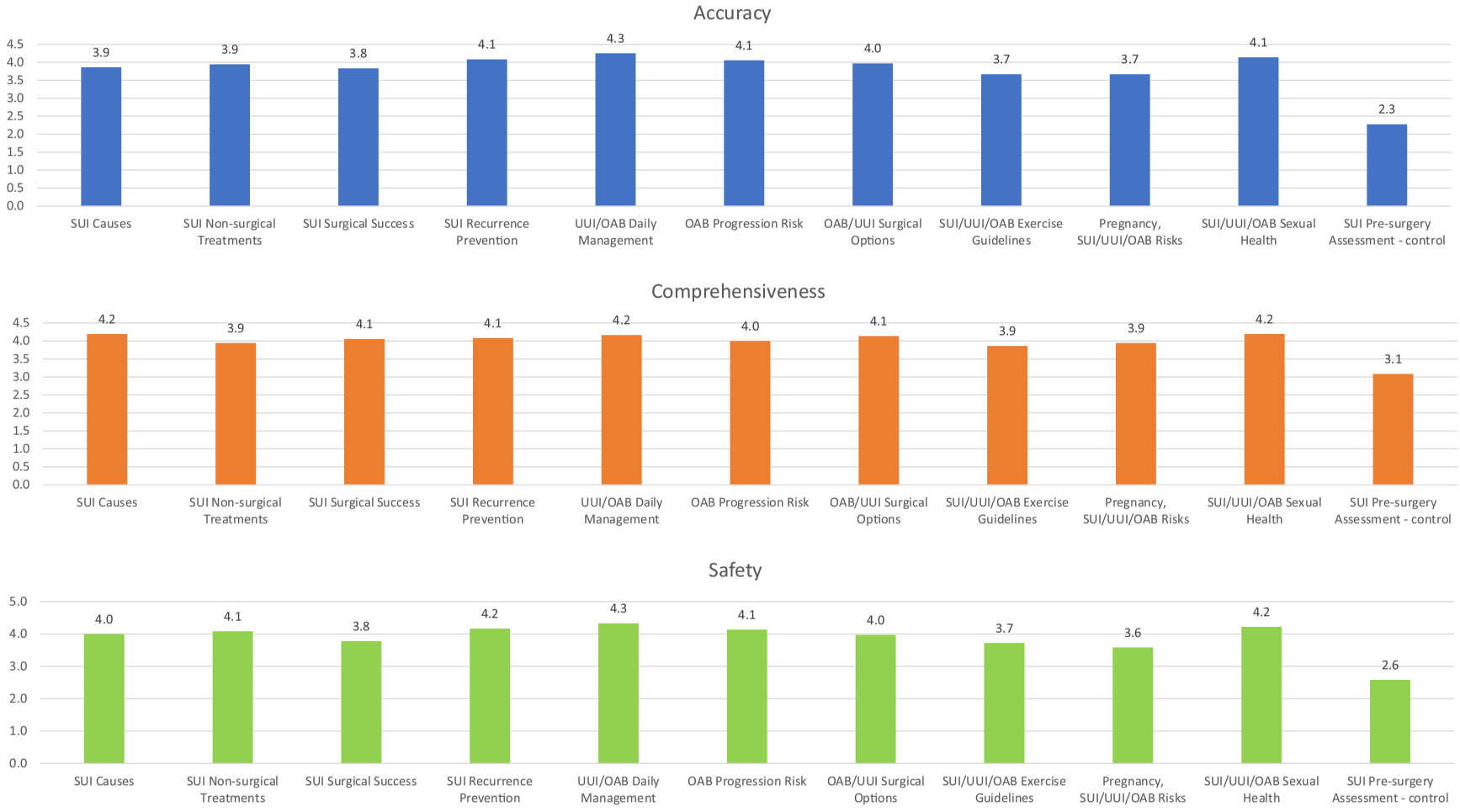


**FIGURE 1** Distribution of ratings based on criteria across ChatGPT's answers.

of doctors.[18,19] This invites deep reflection on our profession's enduring value and our approach to technologic changes.[20] Positioned at this junction, we can either fear AI or embrace it as a beneficial adjunct to traditional medical practices. Our study reveals that, while not perfect, ChatGPT poses no threat to patient safety and can serve as a preliminary resource to in-person evaluations, empowering patients with initial knowledge. The emergence of AI in medicine does not render human interaction and clinical wisdom obsolete. Integrating AI can enhance our care by analyzing data quickly while we provide empathy and nuanced judgment. As we advance with AI in health care, we must uphold ethical standards, protect patient privacy, and maintain the doctor–patient bond. This study highlights the fact that tools like ChatGPT, although imperfect, can enhance patient experience and confidence in the healthcare process.

Our study's strength is evident in its global participation, practical relevance, and the depth of expertise among the respondents. By involving urogynecologists worldwide, we were able to stress the universal relevance of our findings, especially concerning UI. The anonymity regarding ChatGPT's involvement, alongside the use of a control question, strengthened the study's methodologic rigor, ensuring objective and reliable assessments.

Nevertheless, the study faces limitations that merit careful consideration. The evaluation scale we used lacks validation for AI-generated content, raising concerns about whether scores, such as the average of 4.0, accurately reflect quality for patient education. Without comparative benchmarks, the efficacy of this scale is uncertain. Our study did not include comparisons with expert-generated responses, which would have provided a clearer baseline for judging the AI's performance. Future research should incorporate established methods, blind comparisons, and control for evaluator bias to rigorously assess AI-generated answers. Additionally, our methodology involved only expert review, not patient interpretation, which could differ significantly. Conclusions regarding the AI tool's influence on patient decision-making should be approached with caution. The small sample size and variability among responses introduce uncertainty in quantifying ChatGPT's capabilities. We also did not compare ChatGPT with other information sources or AI tools, leaving its relative efficacy unexplored. Lastly, our study used the paid version of ChatGPT, which may differ from the free version, affecting generalizability. AI technologies, including ChatGPT, are continuously updated, meaning that future versions may differ in capabilities, complicating the evaluation of their consistency and reliability. These limitations underscore the need for prudent optimism in considering AI's role in health care—a tool of immense potential that demands judicious application and ongoing evaluation to ensure its alignment with clinical standards and patient care objectives.

In conclusion, our study highlights the potential of AI, specifically ChatGPT, in addressing UI within the field of urogynecology. It demonstrates that while ChatGPT is not flawless, it may provide valuable preliminary insights that could precede traditional patient care. This integration suggests a model of augmented medicine where AI supports both patients and healthcare professionals, allowing for enhanced patient education on UI and enhancing patient confidence. Our findings advocate for the judicious use of AI in medical practice, while refining these tools to maintain their accuracy and reliability. As we embrace digital advancements, our study calls for a balanced approach to integrate AI in health care, ensuring that it aligns with the core values and ethical standards of the medical profession.

## AUTHOR CONTRIBUTIONS

RR, OZ, MR, and AYW designed, planned, and conducted the study, and wrote the manuscript; and OEO and BAO analyzed the data and edited the manuscript.

## CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest.

## DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

## ORCID

*Misgav Rottenstreich* https://orcid.org/0000-0003-4875-1455

## REFERENCES

1. Reynolds WS, Dmochowski RR, Penson DF. Epidemiology of stress urinary incontinence in women. *Curr Urol Rep*. 2011;12:370-376. doi:10.1007/S11934-011-0206-0
2. Game X, Dmochowski R, Robinson D. Mixed urinary incontinence: are there effective treatments? *Neurourol Urodyn*. 2023;42:401-408. doi:10.1002/nau.25065
3. Kenton KS, Smilen SW, ACOG Practice Bulletin No. 155: urinary incontinence in women. *Obstet Gynecol*. 2015;126:E66-E81. doi:10.1097/AOG.0000000000001148
4. Felde G, Engeland A, Hunskaar S. Urinary incontinence associated with anxiety and depression: the impact of psychotropic drugs in a cross-sectional study from the Norwegian HUNT study. *BMC Psychiatry*. 2020;20:521. doi:10.1186/S12888-020-02922-4
5. Schluter PJ, Askew DA, Jamieson HA, Arnold EP. Urinary and fecal incontinence are independently associated with falls risk among older women and men with complex needs: a national population study. *Neurourol Urodyn*. 2020;39:945-953. doi:10.1002/NAU.24266
6. Lane GI, Hagan K, Erekson E, Minassian VA, Grodstein F, Bynum J. Patient-provider discussions about urinary incontinence among older women. *J Gerontol A Biol Sci Med Sci*. 2021;76:463-469. doi:10.1093/GERONA/GLAA107
7. Khera R, Butte AJ, Berkwits M, et al. AI in medicine-JAMA's focus on clinical outcomes, patient-centered care, quality, and equity. *JAMA*. 2023;330:818-820. doi:10.1001/JAMA.2023.15481
8. Lazzerini M, Barbi E, Apicella A, Marchetti F, Cardinale F, Trobia G. Delayed access or provision of care in Italy resulting from fear of COVID-19. *Lancet Child Adolesc Health*. 2020;4:e10-e11. doi:10.1016/S2352-4642(20)30108-5
9. Kuroiwa T, Sarcon A, Ibara T, et al. The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. *J Med Internet Res*. 2023;25:e47621. doi:10.2196/47621
10. Buhr CR, Smith H, Huppertz T, et al. ChatGPT versus consultants: blinded evaluation on answering otorhinolaryngology case-based questions. *JMIR Med Educ*. 2023;9:e49183. doi:10.2196/49183
11. Grünebaum A, Chervenak J, Pollet SL, Katz A, Chervenak FA. The exciting potential for ChatGPT in obstetrics and gynecology. *Am J Obstet Gynecol*. 2023;228:696-705. doi:10.1016/J.AJOG.2023.03.009

12. Introducing ChatGPT n.d. Accessed March 23, 2024. https://openai.com/blog/chatgpt
13. Ozgor BY, Simavi MA. Accuracy and reproducibility of ChatGPT's free version answers about endometriosis. *Int J Gynaecol Obstet*. 2023;165:691-695. doi:10.1002/IJGO.15309
14. Peled T, Sela HY, Weiss A, Grisaru-Granovsky S, Agrawal S, Rottenstreich M. Evaluating the validity of ChatGPT responses on common obstetric issues: potential clinical applications and implications. *Int J Gynaecol Obstet*. 2024. doi:10.1002/IJGO.15501
15. Elenskaia K, Haidvogel K, Heidinger C, Doerfler D, Umek W, Hanzal E. The greatest taboo: urinary incontinence as a source of shame and embarrassment. *Wien Klin Wochenschr*. 2011;123:607-610. doi:10.1007/S00508-011-0013-0
16. Horrocks S, Somerset M, Stoddart H, Peters TJ. What prevents older people from seeking treatment for urinary incontinence? A qualitative exploration of barriers to the use of community continence services. *Fam Pract*. 2004;21:689-696. doi:10.1093/FAMPRA/CMH622
17. Koch LH. Help-seeking behaviors of women with urinary incontinence: an integrative literature review. *J Midwifery Womens Health*. 2006;51:e39-e44. doi:10.1016/J.JMWH.2006.06.004
18. Ahuja AS, Schmidt CE. The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ*. 2019;7:e7702. doi:10.7717/PEERJ.7702
19. Shuaib A, Arian H, Shuaib A. The increasing role of artificial intelligence in health care: will robots replace doctors in the future? *Int J Gen Med*. 2020;13:891-896. doi:10.2147/IJGM.S268093
20. Oh S, Kim JH, Choi SW, Lee HJ, Hong J, Kwon SH. Physician confidence in artificial intelligence: an online mobile survey. *J Med Internet Res*. 2019;21:e12422. doi:10.2196/12422

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Rotem R, Zamstein O, Rottenstreich M, O'Sullivan OE, O'reilly BA, Weintraub AY. The future of patient education: A study on AI-driven responses to urinary incontinence inquiries. *Int J Gynecol Obstet*. 2024;167:1004-1009. doi:10.1002/ijgo.15751