Use of artificial intelligence by patients for understanding of diagnosis of urogynecologic conditions and its role in treatment decisions

Background

Artificial intelligence in medicine and the use of machine learning to improve patient care and outcomes is a quickly developing field.[1] In urogynecology, artificial neural networks have been developed to assess various aspects of care such as pelvic organ prolapse, urodynamic findings, and urinary incontinence, and have been proven to have mixed accuracy for identification and prediction of disease.[2–4]

Interest is additionally building in the use and accuracy of artificial intelligence chatbot programs such as ChatGPT for patient diagnosis and counseling. These programs shown impressive accuracy in the field of obstetrics and gynecology while excluding misinformation; however, there are some misgivings about the frequency of updates to its source data.[5] A recent study of Chat GPT accuracy compared with patient pamphlets about pelvic organ prolapse found comparable accuracy and completeness.[6] Two-way chatbots have additionally started to be implemented for patient use and communication in the reproductive endocrinology and infertility subspecialty.[7] Given the novelty of this field, no current literature exists regarding the use of AI chatbot technology for patient care and patient counseling in urogynecology.

Study Design

This will be a single-center, prospective, randomized, non-blinded study examining patient use of AI Chatbot technology at initial visits to supplement understanding of urogynecologic problems.

Study Aims/Objectives

The primary aim of this study is to investigate the effect of use of an AI Chatbot platform on patient understanding of her disease process and treatment options prior to or following a consult with a urogynecologist at the initial visit. The secondary aims are to evaluate the accuracy of chatbot-provided diagnosis (for participants applicable through randomization) and counseling information, and to evaluate patient satisfaction with their visit.

Specifically, this study will evaluate:

(1) the effect of use of an AI Chatbot platform on patient understanding of her condition prior to or following a consult with her urogynecologist at her initial visit
   a. via patient self-evaluated understanding of her condition
   b. via the Decisional Conflict Scale (DCS) validated in English and Spanish
(2) the accuracy of chatbot-provided diagnosis and counseling information via expert clinician review
(3) patient satisfaction with her visit via patient questionnaire

Hypothesis

In this study, we hypothesize that use of AI Chatbot technology to discuss urogynecologic diagnoses will improve patient understanding of their primary presenting problem.

Human Subjects

Inclusion criteria

- female
- presenting for their initial evaluation by a urogynecology physician for one of the following:

  - urinary incontinence (UI)
  - lower urinary tract symptoms (LUTS)
  - pelvic organ prolapse (POP)
- ≥18 and ≤89 years old
- any race/ethnicity
- able to read or speak English or Spanish
- able/willing to consent to participate

Exclusion criteria

- male
- primary presenting problem other than UI, LUTS, or POP
- non-English or non-Spanish speaking
- pregnant or lactating, as this may affect patient treatment counseling
- unable/unwilling to consent to participate

Procedures

Patients identified on check-in for evaluation any of these three problems as their chief complaint will be asked to participate in this study following check-in for their appointment by dedicated research assistant personnel. If the patient agrees to participate, she will be brought to a private office after completing her check-in paperwork. Research assistant personnel will review check-in paperwork to ensure completion of included PFDI assessment tool. She will then undergo an in-person consent process with the research assistant. If Spanish-speaking, consent will take place with the assistance of a phone-based or in-person interpreter. Demographic information then will be collected. During the consenting process, each individual will be given the choice of contact at three-month follow-up by phone call, text, or e-mail, which she will initial on "contact preferences" section of the informed consent document. She will have the option of completing it over the phone with research assistant personnel, or online via the REDCap survey function.

Participants will be randomized to one of three arms:

1. use of an AI chatbot prior to seeing the urogynecologist
2. use of an AI chatbot following a consult with the urogynecologist
3. no use of an AI chatbot at the time of her visit.

For participants randomized to arm 1

Following consent, the participant will immediately be provided with an iPad with the ChatGPT-4 application already loaded. They will be provided with a brief orientation to the program: "You may ask this program any question you like by typing the question in this text bar and pressing enter, or by pressing this icon and speaking your question into the iPad." They will then be provided with the following instructions: "Please ask the program about the most important problem that brought you in for this visit today." If they struggle with these instructions or need more clarification, the research assistant will then be allowed to provide the following prompts:

1. Ask the chatbot, based on your symptoms at home, what your diagnosis is.
2. Ask the chatbot what the treatment options might be based on the diagnosis you think is most likely.

The participant will be allowed up to five follow-up/clarification entries into the Chat GPT program, but may state she is done asking questions at any time. This should take no more than five minutes of the participant's time. The questions and answers will be downloaded and saved on a secure

hard drive under that participant's unique coded identifier. After completing this, the participant will be returned to the waiting room, and will proceed through her visit as normal. After her visit, she will again be brought to a private office, and will be given a questionnaire (to be completed via the survey application of REDCap) evaluating her knowledge of her condition, medical decision making, and satisfaction with medical visit, counseling, and conclusions on diagnosis. Completion of this questionnaire should take no more than ten minutes of the participant's time.

For participants randomized to arm 2

After being consented, the participant will be returned to the waiting room and will proceed through her visit as normal. After the visit, she will again be brought to a private office. She will be provided with an iPad with the ChatGPT-4 application already loaded. She will be provided with a brief orientation to the program: "You may ask this program any question you like by typing the question in this text bar and pressing enter, or by pressing this icon and speaking your question into the iPad." She then will be provided with the following instructions: "Please ask the program about the most important problem that brought you in for this visit today." If she struggles with these instructions or needs more clarification, the research assistant will then be allowed to provide the following prompts:

1. Ask the chatbot, based on your symptoms at home, what your diagnosis is.
2. Ask the chatbot what the treatment options might be based on the diagnosis you think is most likely.

The participant will be allowed up to five follow-up/clarification entries into the Chat GPT program, but may state that she is done asking questions at any time. This should take no more than five minutes of the participant's time. The questions and answers will be downloaded and saved on a secure hard drive under that participant's unique coded identifier. She will then be given a questionnaire (to be completed via the survey application of REDCap) evaluating her knowledge of her condition, medical decision making, and satisfaction with medical visit, counseling, and conclusions on diagnosis. Completion of this questionnaire should take no more than ten minutes of the patient's time.

For participants randomized to arm 3

After undergoing the consent process, participants will be returned to the waiting room to await the beginning of their appointment. After completing their visit as normal, they will again be brought to a private office. They will be given a questionnaire (to be completed via the survey application of REDCap) evaluating their knowledge of their disease, medical decision making, and satisfaction with medical visit, counseling, and conclusions on diagnosis. Completion of this questionnaire should take no more than ten minutes of the patient's time.

All participants

After three months, participants in all three arms will be asked to repeat a questionnaire identical to the one that they completed at their initial visit, evaluating their knowledge of their condition (individualized to what patient self-identified as their primary problem), satisfaction with their medical visit, and perception of counseling conclusions. Participants will be mailed a $25 gift card for participation after their three-month survey is completed.

Chart review will be conducted examining patient's office consult note to collect primary physician-documented diagnoses, and any treatment options offered/selected. There will be no monitoring of patient electronic record information or clinical outcomes beyond this review of the office consult note. Physicians also will be asked questions to evaluate their perception of patient's primary presenting problem at the initial visit.

Risks/Benefits to Participants

Potential risks and mitigation strategy:

This is a minimal-risk study. The only potential risks are:

(a) time/inconvenience needed to complete the surveys. This risk is mitigated by reviewing the study's consent, in which time/study duration are stated, and answering any questions;

(b) the possibility, inherent with all research studies, of accidental breach of confidentiality. This risk is mitigated as described below in Data Storage/Security. Additionally, Chat GPT4 embedded features will be used to prevent the platform from saving conversations and any resultant machine learning or model training from any patient-entered information for the duration of this study.

Potential benefits:

There may be no direct benefit to/for the participant. There may be a benefit to future patients, as we hope to investigate how to help patients better understand their disease processes, and navigate the quickly developing world of decision-aid technology.

Randomization

A randomization scheme will be programmed in Excel and provided to the research assistant who is doing the consenting. Each participant will be assigned a procedure arm based on the next sequential number; *e.g.*, the study's third participant would be assigned to whatever arm is listed for ID 3. Randomization will be 1:1:1, and three-month attrition is expected to be the same for each arm.

Data Collection/Use

Demographic and clinical data will be collected:

- Demographics
  - Age
  - Race/ethnicity
  - Parity
  - Education level
  - Mailing address (for gift cards)
  - Prior to your visit, did you research your problem?
    - If yes, on what platform(s)?
- Intake questionnaire (PFDI) (already collected by clinic as part of standard intake packet)
- Health literacy assessment (administered by research assistant)
  - Short Assessment of Health Literacy (SAHL) questionnaire (validated in English and Spanish, included with application separately)
- Research assistant questions
  - Did you have to supply supplementary instructions?
  - Did participant speak into program or type?
  - What language did participant use to interact with platform?
- Chat GPT conversation
  - Expert analysis for accuracy and content of Chat GPT conversation at baseline (review by two independent experts in the field, with adjudication by a third for any areas of disagreement. Chat GPT conversations will be screened by study personnel after enrollment of 15 patients in each arm for misleading information or inaccuracy. If

consistent inaccuracies of Chat GPT-provided information (greater than or equal to two of the same/similar inaccuracies on a specific topic) are identified, study personnel will intervene via personal contact of participants and verbal correction of inaccurate information.
        - Five point Likert scale: score of 3="equally accurate", score of 5="much more accurate") (as compared to standard patient information leaflet)
        - PEMAT tool (included with application separately)
        - SMOG readability formula (included with application separately)
    - "Themes" qualitative analysis for patient verbiage
- Office consultation note
    - Primary diagnoses
    - Treatment options offered
    - Treatment option selected
- Physician questionnaire
    - What do you believe is patient's primary complaint or presenting problem? [multiple choice, choose one: uterovaginal prolapse, incontinence, LUTS, not sure (applicable if patient has several problems and they can't tell which one is primary)]
    - Did you feel the patient understood their diagnosis? (non-validated scale, for strongly agree to strongly disagree)
    - Did you feel the patient understood their counseling and options? (non-validated scale, for strongly agree to strongly disagree)
    - Did you feel had adequate previous knowledge about their condition or treatment options? (non-validated scale, for strongly agree to strongly disagree)
- Post-consultation participant questionnaire (immediately post-consultation, and at three months post-consultation)
    - What is your diagnosis for your primary problem? [multiple choice, choose one: uterovaginal prolapse, incontinence, LUTS, not sure]
    - What are possible treatments for your diagnosis? (multiple choice with multiple answer selection, answer choice provided based on above selection)
        - Prolapse – monitoring only, physical therapy, pessary, surgery
        - Incontinence – monitoring only, behavioral modification/conservative treatment, physical therapy, vaginal insert/pessary, medication, bulking, Botox, surgery
        - LUTS – monitoring only, behavioral modification/conservative treatment, medication, surgery
    - What treatment plan did you and your doctor decide upon for your primary problem?
        - Prolapse – monitoring only, physical therapy, pessary, surgery
        - Incontinence – monitoring only, behavioral modification/conservative treatment, physical therapy, vaginal insert/pessary, medication, bulking, botox, surgery
        - LUTS – monitoring only, behavioral modification/conservative treatment, medication, surgery
    - Understanding of diagnosis items (non-validated scale, Likert scale for strongly agree to strongly disagree)
        - I know what my diagnosis is
        - My diagnosis has been clearly described to me
        - I understand what my diagnosis means for my life
    - Decisional Conflict Scale[8] [16 items, validated scale (included with application separately)]

- Satisfaction
  - How satisfied are you with your medical visit? (Likert scale 1-10)
  - How satisfied are you with the counseling your received about your primary problem? (VAS scale 0-10)
  - How satisfied are you about you and your doctor's conclusions about your diagnosis? (VAS scale 0-10)
  - How satisfied are you about you and your doctor's conclusions about your treatment plan? (VAS scale 0-10)
- Chatbot Satisfaction (non-validated Likert scale, for strongly agree to strongly disagree)
  - The chatbot was easy to work with
  - The chatbot identified my diagnosis
  - The chatbot helped me understand my diagnosis
  - The chatbot identified my treatment options
  - The chatbot helped me understand my treatment options
  - The information you received from the chatbot was similar to the information you received from the doctor

Data Storage/Security

Signed informed consents will be stored in a locked file cabinet at Hartford Hospital. Data will be entered into a Hartford HealthCare REDCap database using study IDs for all participants and removing personal identifiers. A separate linked, password-protected Excel file documenting study ID and personal identifiers will be kept on a secure HHC network server. All chatbot questionnaires will be downloaded and saved in PDF format (then converted to Word for editing/analytical processing) with the participant's study ID in a folder on a secure HHC network server, after which they will be deleted from the web platform. No identifying information will be collected in association with chatbot question documents. Only investigators named on the study will have access to the data, and all data will be maintained for six years following study closure.

Power Analysis/Sample Size Estimate

The time frame for the study will dictate the sample size. Factoring in the approximate number of eligible patients (125), an expected participation rate of 80%, and a three-month attrition rate of 10%, the study will comprise approximately 90 sets of evaluate data, with 30 records in each of the three arms.

The DCS will be used as the basis for power. It is a 16-item, Likert scale questionnaire with five subscales. The total score will be used, which will be calculated from a sum of the 16 items (each 0-4), multiplied by 25, to get a range of 0 (no decisional conflict) – 100 (maximum decisional conflict). Scores <25 are associated with implementing decisions; scores >37 are associated with decision delay or being unsure about implementation.[7] Thus, a score of 31 will be used as a midpoint. A one-way analysis of variance (ANOVA) with a sample of 90 divided equally among three arms would afford 90% power to detect a difference, using a non-central F test with an alpha level of 0.05, if the means of each arm under the alternative hypothesis are 38, 31, 24 for chatbot before, chatbot after, and no chatbot, respectively. The means of each arm under the null hypothesis are equal. The common standard deviation (SD) of the responses is assumed to be 15. If the SD were 17, power would be 81%.

Data/Statistical Analyses

Descriptive data will be presented to summarize characteristics of participants in each arm. All continuous data will be evaluated for normality of distribution. Normally distributed data (*e.g.*, age) will be presented as a mean and standard deviation. Non-normally distributed data (*e.g.*, parity) will be

presented with a median and interquartile range. Categorical data (*e.g.*, race/ethnicity) will be presented as a frequency, using percentage.

Inferential statistics will be used to compare differences, if any, between the three arms. Continuous data, including DCS scores, will be compared with an ANOVA. If a statistically significant result is found, an appropriate *post hoc* test (*e.g.*, Scheffé's multiple range test or Tukey's honestly significant difference) will be used to evaluate whether there is a difference between any two-arm comparisons. For non-normally distributed data, a Kruskal-Wallis test will be used. A chi square test will be used for categorical comparisons, although the sample size (or cell sizes) may be limiting. If observed frequencies are sufficiently small, a *post hoc* power analysis will be run.

Data will be compared by patient-identified diagnosis within each arm, which is expected to be equally distributed (*i.e.*, 10/arm). However, most of these results will be descriptive, as there likely will be insufficient power to detect differences. Patient-identified diagnosis also will be compared with physician-identified diagnosis for agreement. A kappa (κ) or weighted kappa (κ_w) statistic, depending on distribution, will be generated for each arm and strength of agreement will be compared.

Chatbot-provided information collected for each participant will be evaluated and scored by two independent expert clinicians for accuracy and appropriateness on a scale of 1-10, with any score differing among the clinical experts by more than 2 points arbitrated by a third reviewer.

Participant questionnaires will be compared between arms for patient retention of appropriate medical information as well as patient satisfaction with diagnosis and counseling. Three-month questionnaires will be analyzed for retention of information and satisfaction, using an ANOVA based on comparison of paired differences.

All quantitative analyses will be conducted with SPSS v. 29 (IBM; Armonk, NY 2022). Qualitative analyses will be conducted with Word (Microsoft, 2016) or nVivo (QSR International, 2021). An *a priori* alpha level of 0.05 will be used such that all quantitative results yielding $p<0.05$ will be deemed statistically significant. No alpha level is applicable for the qualitative findings, since only themes will be elucidated.

Significance to HH and Its Patients

The existing literature lacks data regarding patient use of technology aids such as AI chatbot technology in medical decision making, specifically in the area of urogynecology. This study will allow for evaluation of patient's use of one such tool, in addition to the accuracy of information provided. This will be beneficial as it will improve physician understanding of, and counseling about, these tools for HH patients in the future.

References

1. Daykan Y, O'Reilly BA. The role of artificial intelligence in the future of urogynecology. *Int Urogynecol J*. 2023;34(8):1663-1666. doi:10.1007/s00192-023-05612-3
2. Robinson CJ, Swift S, Johnson DD, Almeida JS. Prediction of pelvic organ prolapse using an artificial neural network. *Am J Obstet Gynecol*. 2008;199(2):193.e1-193.e6. doi:10.1016/j.ajog.2008.04.029
3. Serati M, Salvatore S, Siesto G, et al. Urinary symptoms and urodynamic findings in women with pelvic organ prolapse: Is there a correlation? results of an artificial neural network analysis. *Eur Urol*. 2011;60(2):253-260. doi:10.1016/j.eururo.2011.03.010
4. Qureshi A, Mathur A, Alshiek J, Shobeiri SA, Wei Q. Utilization of Artificial Intelligence for Diagnosis and Management of Urinary Incontinence in Women Residing in Areas with Low Resources: An Overview. *Open J Obstet Gynecol*. 2021;11(04):403-418. doi:10.4236/ojog.2021.114040

5. 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(6):696-705. doi:10.1016/j.ajog.2023.03.009

6. Johnson CM, Bradley CS, Kenne KA, et al. Evaluation of ChatGPT for Pelvic Floor Surgery Counseling. *Urogynecology*. 2024;30(3):245-250. doi:10.1097/SPV.0000000000001459

7. Acker A, Senapati S, Dokras A. Barriers to access: findings from an implementation study of an artificial intelligence–augmented 2-way chatbot for fertility care. *Fertil Steril*. 2023;120(1):199-201. doi:10.1016/j.fertnstert.2023.04.016

8. https://decisionaid.ohri.ca/docs/develop/User_Manuals/UM_Decisional_Conflict.pdf (last accessed 2/2/24)