Is Artificial Intelligence-Assisted Pregnancy Counseling Feasible? An Evaluation of the Quality of ChatGPT Responses

Mücahit Furkan Balcı[1], Celal Akdemir[2], Fatih Yıldırım[3]

1 Department of Obstetrics and Gynecology, Torbalı State Hospital, İzmir, Türkiye
2 Department of Gynecologic Oncology, İzmir City Hospital, İzmir, Türkiye
3 Department of Obstetrics and Gynecology, Tepecik Training and Research Hospital, İzmir, Türkiye

Abstract

Aim: To evaluate the quality of ChatGPT-generated responses to commonly asked questions during pregnancy, based on expert assessments using predefined criteria: accuracy, completeness, and safety.

Methods: A total of 20 board-certified obstetricians evaluated 15 ChatGPT-generated responses to frequently encountered pregnancy-related questions. Each response was assessed using a 5-point Likert scale across three domains. Statistical comparisons were conducted to evaluate differences among criteria and question categories.

Results: ChatGPT received an overall mean score of 4.1 across all criteria. Accuracy was the highest-rated criterion (mean 4.27 ± 0.31), followed by completeness (3.85 ± 0.30) and safety (3.78 ± 0.36) (P = 0.019). Responses to general knowledge questions scored significantly higher than those related to symptoms or follow-up guidance (P = 0.041). The most favorably rated response pertained to sleep positions during pregnancy (mean 4.5), while painkiller safety scored the lowest (mean 3.5).

Conclusions: ChatGPT demonstrates strong potential in delivering accurate and comprehensible pregnancy-related information. However, its limitations in clinical safety and completeness—particularly in symptom-related topics—suggest that it should be used as an adjunct to, not a replacement for, professional medical guidance. Further validation across diverse clinical scenarios and standardized evaluation tools is necessary to ensure safe integration into patient education.

Keywords: *ChatGPT; artificial intelligence; pregnancy; obstetrics; medical safety; expert evaluation*

1. Introduction

Artificial intelligence (AI) tools such as ChatGPT have undergone accelerated adoption in medical settings because of their facility for producing syntactically coherent and contextually salient outputs in response to user prompts.[1] While systematic syntheses document value for patient education, research support, and certain aspects of clinical practice, these advantages are tempered by persistent shortcomings in factual reliability and by ethical liabilities related to bias, accountability, transparency, and the protection of sensitive health data.

There is growing evidence that ChatGPT responses to patient queries are often rated higher than clinician responses in terms of empathy and content quality. For example, Ayers et al. (2023) reported that in an analysis of 195 patient questions retrieved from Reddit, ChatGPT responses were rated as higher in quality and empathy than physician answers in approximately 79% of cases.[2] However, ChatGPT's performance in highly specialized fields such as in vitro fertilization, obstetrics, and gynecology remains underexplored. A recent study published by Karger indicated that the model offers limited clinical accuracy when responding to obstetric questions. Nevertheless, it was also reported that ChatGPT-4 achieved significantly higher success rates on obstetrics and gynecology board certification exams compared to earlier versions.[3]

Pregnancy is a sensitive period during which expectant mothers frequently seek information, both for medical and emotional reassurance.[4-6] Therefore, evaluating the reliability, comprehensiveness, and safety of ChatGPT responses to common pregnancy-related questions is of considerable importance in clinical practice.

The aim of this study is to assess ChatGPT (GPT-4.0) responses to 15 frequently asked pregnancy-related questions, based on evaluations by board-certified obstetricians and gynecologists across three domains: accuracy, completeness, and safety. This assessment will help explore the potential role of AI-supported consultation

Corresponding Author: Mücahit Furkan Balcı, mucahitfurkanbalci@gmail.com, *Received: 30.07.2025, Accepted: 26.09.2025, Available Online Date: 30.09.2025* *Sağlam Apartment, Apt. 6, No. 15, 6138 Street, Dedebaşı Neighborhood, Karşıyaka, İzmir, Türkiye.* *https://doi.org/10.36516/jocass.1753846*

tools in enhancing information sharing during pregnancy.

2. Materials and Methods

2.1. Study Design and Setting

In this study, a dataset comprising 15 questions was developed to evaluate the validity of artificial intelligence-generated responses to frequently asked pregnancy-related inquiries. The questions were designed to reflect common information-seeking behaviors of pregnant individuals based on clinical practice. Particular attention was given to ensuring that the questions addressed current, accurate, and guidance-oriented content relevant to pregnancy.

Two researchers specialized in obstetrics and gynecology collaboratively created the question set to ensure both clinical and face validity. The final list of questions was formulated in Turkish and structured to be easily understood by the general population.

2.2. Use of ChatGPT and Collection of Responses

The finalized 15 questions were submitted to ChatGPT (GPT-4.0; OpenAI), using its most up-to-date version as of March 2024. Data entry was performed by an independent researcher who was not involved in the study design, thereby ensuring objectivity. Each question was entered into the platform using a standardized prompt: *"I am a pregnant woman. Please provide a short, clear, and understandable answer."* All responses generated by ChatGPT were collected in Turkish and included directly in the evaluation process.

2.3. Collection Expert Evaluation and Assessment Criteria

To assess the quality of ChatGPT's responses, 30 obstetrics and gynecology specialists working across Türkiye were invited to participate in the study. The first 20 respondents were included in the final sample.

Each participant was asked to evaluate the ChatGPT responses based on the following three criteria:

1. Accuracy: Scientific and clinical correctness of the response
2. Comprehensiveness: The extent to which the response sufficiently addressed all aspects of the question
3. Safety: Whether the response contained any potentially harmful information for the mother or fetus

Each criterion was rated using a 5-point Likert scale (1 = Strongly disagree; 5 = Strongly agree). The evaluation form was prepared in Turkish and administered online.

2.4. Data Analysis

Mean scores were calculated for each evaluation criterion. The proportion of responses receiving a score of ≥4 was defined as a "positive rating." A threshold score of 4 was adapted from similar studies in obstetrics and gynecology literature. Performance scores were further analyzed based on question type (e.g., treatment-related, diagnostic, emergency). One-way ANOVA and post-hoc Tukey tests were applied to determine statistically significant differences among criteria and question categories.

3. Results

Of the 30 obstetricians and gynecologists invited to participate, 20 completed the questionnaire and were included in the evaluation, resulting in a response rate of 66.7%. The median professional experience of the participants was 6.2 years (IQR: 3–11 years).

Table 1 presents the overall evaluation results of ChatGPT-generated responses. When all questions and criteria were assessed collectively, the mean total score was 4.1, and 76.3% of the responses received a rating of ≥4.

Table 1

Overall Evaluation of ChatGPT Responses

Evaluation Criterion	Mean Score (1-5)	Proportion Rated ≥4
All responses (overall)	4.1	76.3%
Accuracy	4.3	81.0%
Comprehensiveness	3.9	69.2%
Safety	3.8	67.5%

Among the 15 ChatGPT responses evaluated, the response regarding sleep positions during pregnancy received the highest average score (4.5), with 95% of experts rating it ≥4. In contrast, the response on the use of over-the-counter painkillers during pregnancy had the lowest average score (3.5), and only 48% of experts rated it ≥4.

Table 2

Distribution of Evaluation Scores by Topic

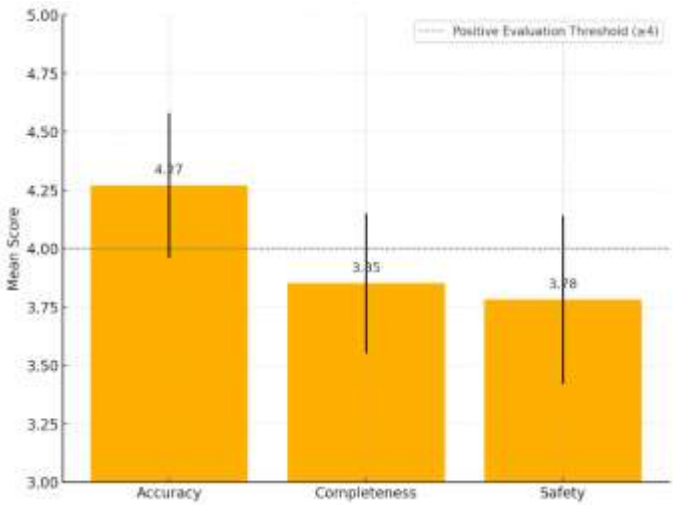
Pregnancy Topic	Mean Score(1-5)	Proportion Rated ≥4
Sleep position	4.5	95%
Early pregnancy symptoms	4.2	90%
Ultrasound safety	4.3	86%
First prenatal visit	4.1	88%
Headache	4.1	84%
Fetal movements	4.0	82%
Exercise	4.0	79%
Sexual intercourse	4.0	76%
Air travel	3.9	70%
Mode of delivery	3.8	66%
Vaginal bleeding	3.7	61%
Contractions	3.6	63%
Abdominal pain	3.7	60%
Water breaking (membrane rupture)	3.5	58%
Painkiller safety	3.5	48%

As shown in Figure 1, the mean scores for all responses exceeded 3.0, with the highest ratings consistently observed under the "accuracy" criterion. While 11 of the 15 responses scored an average of ≥4 in terms of accuracy, fewer responses achieved this threshold for the comprehensiveness and safety criteria.

Among the 15 pregnancy-related topics evaluated by experts, "Sleep Position," "Basic Symptoms," and "Ultrasound Safety" received the highest mean scores across all three evaluation criteria: accuracy, comprehensiveness, and safety. In contrast, topics such as "Painkiller Safety," "Water Breaking," and "Abdominal Pain" were rated lower, particularly in terms of comprehensiveness y and safety.

Figure 1

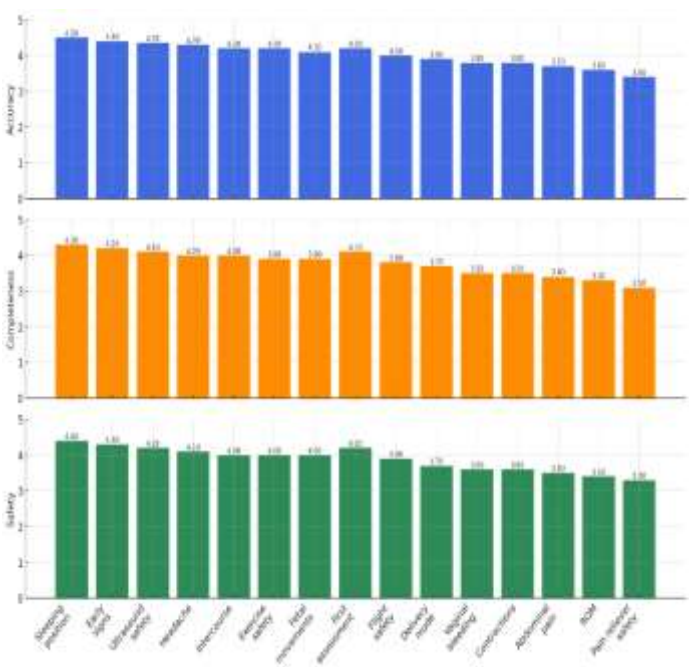
ChatGPT Performance: Mean Evaluation by Criterion



The highest mean score for accuracy was observed in the "Sleep Position" topic (4.5), while the lowest was noted in "Painkiller Safety" (3.4). Scores for comprehensiveness ranged from 4.3 to 3.1, with "Painkiller Safety" again receiving the lowest average in this domain. A similar distribution was observed in the safety evaluations: "Sleep Position" and "Basic Symptoms" exceeded a mean score of 4.3, whereas "Painkiller Safety" and "Water Breaking" fell below 3.5. Figure 2 illustrates the expert ratings of ChatGPT responses by topic and evaluation criterion.

Figure 2

Expert Ratings of ChatGPT Responses



In criterion-based comparisons, accuracy scores were found to be statistically significantly higher than those of the other two criteria (accuracy: 4.27 ± 0.31; comprehensiveness: 3.85 ± 0.30; safety: 3.78 ± 0.36; $P = 0.019$).

Table 3

Distribution of ChatGPT Performance by Evaluation Criterion

Criterion	Mean ± SD	P value
Accuracy	4.27 ± 0.31	
Comprehensiveness	3.85 ± 0.30	0.019
Safety	3.78 ± 0.36	

When the questions were categorized based on their content into general information, follow-up, and symptom-related groups, responses to general information questions received significantly higher total performance scores (4.09 ± 0.34 vs. 4.02 ± 0.29 vs. 3.75 ± 0.33; $P = 0.041$). However, this difference did not remain statistically significant when sub-criteria (accuracy, comprehensiveness, safety) were analyzed separately.

Table 4

ChatGPT Performance by Question Category

Category	Total Score ± SD	Accuracy	Comprehensiveness	Safety	P
General information	4.09 ± 0.34	4.25	3.92	3.95	
Follow-up	4.02 ± 0.29	4.20	3.88	3.90	0.041
Symptom-focused	3.75 ± 0.33	4.00	3.60	3.64	

4. Discussion

4.1. Interpretation of Principal Findings

This study evaluated the quality of ChatGPT-generated answers to common pregnancy-related questions, using expert evaluations based on three key criteria: accuracy, completeness, and safety. The findings revealed generally high ratings, with accuracy scoring the highest among the three. Responses to general knowledge questions

outperformed those focused on symptoms or clinical follow-up, indicating ChatGPT's relative strength in educational rather than clinical contexts. This is consistent with prior AI research suggesting that large language models (LLMs) perform better in factual recall than in clinical reasoning.[1,2]

Among individual topics, answers about sleep positions in pregnancy were rated most favorably, while topics involving medication safety, rupture of membranes, or abdominal pain scored lower. This aligns with Peled et al. [7], who showed that ChatGPT performed better in general health information but struggled in symptom-based or risk-sensitive queries.

4.2. Comparison with Prior Literature

ChatGPT's performance in this study mirrors findings from other fields of medicine. For example, Kung et al. [8] and Gilson et al. [9] demonstrated ChatGPT's ability to pass the United States Medical Licensing Examination (USMLE), suggesting competency in foundational medical knowledge. Likewise, Ayers et al. [2] showed that ChatGPT responses on social media health forums were often preferred over those by physicians, highlighting public trust but also potential risks in misinformation.

Several studies emphasized ChatGPT's ability to synthesize complex material into simplified language. Jeblick et al. [10] showed this in radiology, while Sinha et al. [11] and Das et al. [12] confirmed moderate success in pathology and microbiology. However, Johnson et al. [13] warned of significant variability in output quality, a concern echoed in our findings.

In obstetrics specifically, Grünebaum et al. [14] and Dhombres et al. [15]proposed that ChatGPT could be used to support patient education. Yet, the risk of misguidance in acute clinical scenarios remains. For instance, symptom-related responses in our study (e.g., regarding contractions or bleeding) received lower safety ratings—suggesting that AI should never substitute professional care.

Rahsepar et al. [16] compared ChatGPT with Google Bard in oncology and observed notable variability between platforms, reinforcing our view that future research must assess the consistency and validity of various AI systems in obstetrics.

4.3. Clinical and Ethical Implications

The ethical landscape around AI in healthcare remains complex. Issues of credibility, source transparency, and liability have been raised across recent literature [17-19]. Flanagin et al. [18] and da Silva [19] argue against granting authorship to nonhuman agents, while Hill-Yardin et al. [20] emphasize the risk of academic integrity breaches. These concerns are magnified in maternal care, where patient safety is paramount.

Beyond ethics, social and behavioral aspects must also be considered. Research suggests that many women avoid clinical care due to stigma or embarrassment, particularly with topics like urinary incontinence [21-23]. AI may serve as a low-barrier tool to initiate help-seeking, as shown by Elenskaia et al. [21]. Similarly, Horrocks et al. [22] and Koch [23] emphasize that anonymous platforms could improve access for hesitant populations.

However, the assumption that AI is a safe alternative is misleading. Thirunavukarasu et al. [24] revealed inconsistencies in ChatGPT's clinical decision-making in general practice. Moreover, delayed care has real consequences, as shown by Lazzerini et al. [25], who noted worsened outcomes during COVID-19 due to fear-based care avoidance. ChatGPT, while helpful, must not reinforce such patterns unintentionally.

On the positive side, AI may assist overburdened clinicians. Studies have reported increasing physician confidence in AI-assisted support tools [26], and thought leaders envision an augmented care model where AI complements, rather than replaces, medical professionals [27]. Nonetheless, AI should only serve as an initial step, not a final answer.

4.4. Strengths and Limitations

The strength of this study lies in its use of expert review based on structured scoring across three key dimensions. The panel of 20 obstetricians with a median 6.2 years of clinical experience ensured relevant and diverse insights. Our inclusion of general, symptom-focused, and follow-up questions added depth to the evaluation.

Limitations include the lack of a validated AI-assessment tool and potential evaluator bias due to awareness that responses came from ChatGPT. Furthermore, only one time point and one AI version were analyzed. Johnson et al. [13] noted that ChatGPT may yield different answers even to the same prompt, an inconsistency that our methodology could not capture.

4.5. Future Directions

Future research should use blinded comparisons between AI and expert-generated answers, include patient evaluations, and analyze performance across multiple platforms and languages. As Rahsepar et al. [16] recommended, systematic benchmarking is key. Developing a validated, standardized tool to assess AI output quality in healthcare contexts—especially for safety is a pressing need.

5. Conclusion

This study presents a systematic expert-based analysis aimed at evaluating the accuracy, comprehensiveness, and safety of ChatGPT's responses to frequently asked questions related to pregnancy. The findings indicate that ChatGPT is capable of providing highly accurate and reliable answers, particularly in areas involving general medical information. However, the comprehensiveness and safety of responses were found to be lower in symptom-based inquiries and topics requiring clinical intervention. This highlights the need for caution in using AI-generated content as a source for direct clinical decision-making.

Artificial intelligence demonstrates significant potential in enhancing access to health information and supporting patient education. Nevertheless, to ensure the safe and ethical use of such technologies, rigorous validation studies are necessary, along with the development of standardized assessment tools. Models like ChatGPT may contribute meaningfully to healthcare delivery when applied within clearly defined boundaries and under appropriate oversight; however, they are not yet capable of replacing human expertise.

Statement of ethics

This study did not involve human participants or personal health data; therefore, institutional ethics committee approval was not required. All responses were anonymized, and efforts were made to minimize the dissemination of potentially misleading medical content. None of the reviewers had any affiliation with the ChatGPT development team.

Funding

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

Conflict of interest statement

The authors declare that they have no conflict of interest.

Availability of data and materials

No new data were created or analyzed in this study. Therefore, data sharing is not applicable to this article.

Author contributions

Mücahit Furkan Balcı and Celal Akdemir contributed to the study conception and design. Fatih Yıldırım were responsible for data collection and evaluation. Celal Akdemir conducted the statistical analyses. All authors contributed to drafting the manuscript and approved the final version.

References

1.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. [Crossref]
2.Ayers JW, Poliak A, Dredze M, 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):589–96. [Crossref]
3.Ługowski F, Babińska J, Ludwin A, Stanirowski PJ. Comparative analysis of ChatGPT 3.5 and ChatGPT 4 obstetric and gynecological knowledge. Sci Rep. 2025;15(1):21133. [Crossref]
4.Yeşilçinar İ, Güvenç G, Kinci MF, Bektaş Pardes B, Kök G, Sivaslioğlu AA. Knowledge, fear, and anxiety levels among pregnant women during the COVID-19 pandemic: a cross-sectional study. Clin Nurs Res. 2022;31(4):758–65. [Crossref]
5.Yeşilçinar İ, Kinci MF, Ünver HC, Sivaslioğlu AA. Pregnancy-related anxiety and prenatal attachment in pregnant women with preeclampsia and/or gestational diabetes mellitus: a cross-sectional study. J Clin Obstet Gynecol. 2023;33(1):27–35. [Crossref]
6.Soma-Pillay P, Nelson-Piercy C, Tolppanen H, Mebazaa A. Physiological changes in pregnancy. Cardiovasc J Afr. 2016;27(2):89–94. [Crossref]
7.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 Gynecol Obstet. 2024;166(3):1127–33. [Crossref]
8.Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. [Crossref]
9.Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:e45312. [Crossref]
10.Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol. 2024;34(5):2817–25. [Crossref]
11.Sinha RK, Deb Roy A, Kumar N, Mondal H. Applicability of ChatGPT in assisting to solve higher order problems in pathology. Cureus. 2023;15(2):e35237. [Crossref]
12.Das D, Kumar N, Longjam LA, et al. Assessing the capability of ChatGPT in answering first-and second-order knowledge questions on microbiology as per competency-based medical education curriculum. Cureus. 2023;15(3)
13.Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Research square. 2023:rs. 3. rs-2566942.
14.Grünebaum A, Chervenak J, Pollet SL, Katz A, Chervenak FA. The exciting potential for ChatGPT in obstetrics and gynecology. American Journal of Obstetrics and Gynecology. 2023;228(6):696-705.
15.Dhombres F, Bonnard J, Bailly K, Maurice P, Papageorghiou AT, Jouannic JM. Contributions of Artificial Intelligence Reported in Obstetrics and Gynecology Journals: Systematic Review. J Med Internet Res. Apr 20 2022;24(4):e35465. doi:10.2196/35465
16.Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. Radiology. 2023;307(5):e230922. [Crossref]
17.Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595. [Crossref]
18.Flanagin A, Bibbins-Domingo K, Berkwits M, Christiansen SL. Nonhuman "authors" and implications for the integrity of scientific publication and medical knowledge. JAMA. 2023;329(8):637–9. [Crossref]
19.da Silva JAT. Is ChatGPT a valid author? Nurse Educ Pract. 2023;68:103600. [Crossref]
20.Hill-Yardin EL, Hutchinson MR, Laycock R, Spencer SJ. A chat(GPT) about the future of scientific publishing. Brain Behav Immun. 2023;110:152–4. [Crossref]
21.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(19-20):607–10. [Crossref]
22.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(6):689–96. [Crossref]
23.Koch LH. Help-seeking behaviors of women with urinary incontinence: an integrative literature review. J Midwifery Womens Health. 2006;51(6):e39–44. [Crossref]
24.Thirunavukarasu AJ, Hassan R, Mahmood S, et al. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. JMIR Med Educ. 2023;9:e46599. [Crossref]
25.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(5):e10–1. [Crossref]
26.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(3):e12422. [Crossref]
27.Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ. 2019;7:e7702. [Crossref]