Journal of Surgery and Medicine
e-ISSN: 2602-2079
https://jsurgmed.com/

Performance of artificial intelligence chatbot as a source of patient information on anti-rheumatic drug use in pregnancy

Nurdan Oruçoğlu, Elif Altunel Kılınç

Department of Rheumatology, Mersin University Faculty of Medicine, Mersin, Turkey

ORCID ID of the author(s)

NO: 0000-0002-8613-5373
EAK: 0000-0003-2501-2473

Corresponding Author
Nurdan Orucoglu
Department of Internal Medicine, Division of Rheumatology, Mersin University Faculty of Medicine, Mersin, TR-33343, Turkey
E-mail: nurdanorucoglu@yahoo.com

Ethics Committee Approval
This study did not require ethics committee approval because neither human nor animal data were used.

Conflict of Interest
No conflict of interest was declared by the authors.

Financial Disclosure
The authors declared that this study has received no financial support.

Published
2023 October 4

Abstract

Background/Aim: Women with rheumatic and musculoskeletal disorders often discontinue using their medications prior to conception or during the few early weeks of pregnancy because drug use during pregnancy frequently results in anxiety. Pregnant women have reported seeking out health-related information from a variety of sources, particularly the Internet, in an attempt to ease their concerns about the use of such medications during pregnancy. The objective of this study was to evaluate the accuracy and completeness of health-related information concerning the use of anti-rheumatic medications during pregnancy as provided by Open Artificial Intelligence (AI's) Chat Generative Pre-trained Transformer (ChatGPT) versions 3.5 and 4, which are widely known AI tools.

Methods: In this prospective cross-sectional study, the performances of OpenAI's ChatGPT versions 3.5 and 4 were assessed regarding health information concerning anti-rheumatic drugs during pregnancy using the 2016 European Union of Associations for Rheumatology (EULAR) guidelines as a reference. Fourteen queries from the guidelines were entered into both AI models. Responses were evaluated independently and rated by two evaluators using a predefined 6-point Likert-like scale (1 – completely incorrect to 6 – completely correct) and for completeness using a 3-point Likert-like scale (1 – incomplete to 3 – complete). Inter-rater reliability was evaluated using Cohen’s kappa statistic, and the differences in scores across ChatGPT versions were compared using the Mann–Whitney U test.

Results: No statistically significant difference between the mean accuracy scores of GPT versions 3.5 and 4 (5 [1.17] versus 5.07 [1.26]; *P*=0.769), indicating the resulting scores were between nearly all accurate and correct for both models. Additionally, no statistically significant difference in the mean completeness scores of GPT 3.5 and GPT 4 (2.5 [0.51] vs 2.64 [0.49], *P*=0.541**)** was found**,** indicating scores between adequate and comprehensive for both models. Both models had similar total mean accuracy and completeness scores (3.75 [1.55] versus 3.86 [1.57]; *P*=0.717). In the GPT 3.5 model, hydroxychloroquine and Leflunomide received the highest full scores for both accuracy and completeness, while methotrexate, Sulfasalazine, Cyclophosphamide, Mycophenolate mofetil, and Tofacitinib received the highest total scores in the GPT 4 model. Nevertheless, for both models, one of the 14 drugs was scored as more incorrect than correct.

Conclusions: When considering the safety and compatibility of anti-rheumatic medications during pregnancy, both ChatGPT versions 3.5 and 4 demonstrated satisfactory accuracy and completeness. On the other hand, the research revealed that the responses generated by ChatGPT also contained inaccurate information. Despite its good performance, ChatGPT should not be used as a standalone tool to make decisions about taking medications during pregnancy due to this AI tool’s limitations.

Keywords: anti-rheumatic drugs, artificial intelligence, ChatGPT, pregnancy

How to cite: Oruçoğlu N, Altunel Kılınç E. Performance of artificial intelligence chatbot as a source of patient information on anti-rheumatic drug use in pregnancy. J Surg Med. 2023;7(10):651-655.

Introduction

A significant number of individuals with rheumatic disorders (RMD) receive their diagnoses during the reproductive stages of their lives [1]. Drug usage during pregnancy can frequently cause anxiety; thus, many women with RMDs discontinue medications before pregnancy or during the early stages of their pregnancies [2]. Pregnant women tend to seek health-related information from a variety of sources, as their information needs increase during pregnancy [3]. It has been determined that pregnant women utilize the Internet as a source of information concerning their pregnancies and medications more frequently than they consult medical professionals [4]. Incorrect information obtained from the Internet can increase the tendency of highly worried pregnant women to discontinue their medicine thus leading to exacerbation of the disorder and an increase in the risk of pregnancy-related problems [5]. For this reason, it is very important for patients to have access to accurate information sources during pregnancy. However, some doubts regarding the accuracy and quality of health-related content on the internet exist.

In recent years, the area of computer science known as artificial intelligence (AI) has exhibited substantial development. The language-learning model (LLM) is a natural language processing artificial intelligence (AI) tool that is trained on excessive amounts of datasets and is capable of understanding and generating human-like responses [6]. LLMs and the OpenAI tool "Chat Generative Pre-trained Transformer," or "ChatGPT," in particular, have attracted great interest in medical science lately due to their high performance. One of the most popular, ChatGPT, is a natural language processing (NLP) system developed by OpenAI (OpenAI, L.L.C., San Francisco, CA, USA). Currently, two versions are available: (1) GPT-3.5, which is the fastest and is free to use and (2) GPT-4.0, which has a fee but is regarded as the most powerful version [7]. This version offers several advantages, including more efficiency, higher precision, and cost savings. It has several difficulties, however, including safety issues and limited performance [8]. Furthermore, insufficient evidence concerning the accuracy, reliability, and quality of medical information provided by Chatbots is available.

This study aimed to evaluate the accuracy and completeness of chatbots, including ChatGPT versions GPT 3.5 (free) and 4 (fee for use) in the framework of digital health-related information regarding the use of anti-rheumatic drugs before and during pregnancy.

Materials and methods

In this prospective, cross-sectional study, OpenAI's chatbots ChatGPT versions 3.5 and 4 were used to evaluate the performance of the LLM-based AI for health-related information concerning anti-rheumatic drug use during pregnancy. The reference source for this study was the 2016 European Union of Associations for Rheumatology (EULAR) guidelines entitled “The EULAR points to consider for use of antirheumatic medicines before pregnancy and during pregnancy” [9]. Fourteen queries in this guideline containing information about expert opinions on the use of non-steroidal anti-inflammatory drugs (NSAIDs) and immunosuppressive drugs during pregnancy were used to generate responses.

On September 16, 2023, all domain items were entered as questions into two versions of OpenAI models (GPT-3.5 and -4, August version). English was used as the chat language. Responses obtained from each AI model were analyzed separately by two rheumatology specialists. A single rater submitted questions to the AI programs and recorded the answers. To reduce bias, the other rater had no information about which AI programs generated the answers. In case of disagreement between the scores presented by the raters, the answer was reviewed, and the decision was made by consensus. This final score was utilized for the analysis.

Johnson et al.’s [10] scoring system, which was determined for the ChatGPT study, was used for the accuracy and completeness of the content. The rating of accuracy for each response was assessed using a six-point Likert scale: (1) completely incorrect, (2) more incorrect than correct, (3) Approximately equal correct and incorrect, (4) more correct than incorrect, (5) nearly all correct, and (6) correct.

The completeness scale is based on a 3-point Likert scale: (1) incomplete, missing essential details or information, only partially answering the question; (2) adequate, covering all bases and providing the minimum amount of information required to be considered complete; and (3) comprehensive, covers all areas of the query, and offers more details than what was expected.

The study did not require ethical approval as it did not involve human or animal participants.

Statistical analysis

Data analysis was performed using SPSS software (IBM SPSS Statistics v. 22.0 for Windows; Armonk, NY: IBM Corp). Numbers, percentages, and median (interquartile range) values were used to represent descriptive data. The Shapiro–Wilk test was used to determine normality of the data. Inter-rater reliability and overall agreement between raters were assessed using Cohen’s kappa statistic. According to intra-class correlation coefficient results, positive values ranging from 0 to 0.2 indicated poor agreement, 0.2 to 0.4 indicated fair agreement, 0.4 to 0.6 indicated moderate agreement, 0.6 to 0.8 indicated good agreement, and 0.8 to 1 indicated very good agreement. Differences observed in the scores across ChatGPT versions were compared using the Mann–Whitney U test. Significance was evaluated at the level of P <0.05.

Results

The mean accuracy scores for GPT 3.5 were 5 (1.17) and 5.07 (1.26) for GPT 4 with no statistically significant difference between scores (P=0.769). The mean completeness scores for GPT 3.5 were 2.5 (0.51) and 2.64 (0.49) for GPT 4 with no statistically significant difference between the two versions (P=0.541). Both models had similar mean accuracy and completeness scores (3.75 [1.55] versus 3.86 [1.57)]; P=0.717). Table 1 presents the accuracy and completeness scores regarding the medicine.

Table 1: Accuracy and completeness of responses generated by Chat Generative Pre-trained Transformer (ChatGPT) versions 3.5 and 4 to the questions regarding the use of NSAIDs, synthetic DMARDs, and immunosuppressive medicines in pregnancy

		GPT 3.5	GPT 4	*P*-value
Methotrexate	Accuracy Completeness	6 2	6 3	
Leflunomide	Accuracy Completeness	6 3	6 3	
Sulfasalazine	Accuracy Completeness	6 2	6 3	
Hydroxychloroquine	Accuracy Completeness	6 3	6 3	
Azathioprine	Accuracy Completeness	2 2	4 2	
Cyclophosphamide	Accuracy Completeness	6 2	6 3	
Ciclosporin	Accuracy Completeness	4 2	2 2	
Mycophenolate mofetil	Accuracy Completeness	6 2	6 3	
Prednisone	Accuracy Completeness	4 3	4 3	
NSAIDs	Accuracy Completeness	4 3	4 3	
Colchicine	Accuracy Completeness	5 3	4 2	
Tofacitinib	Accuracy Completeness	5 3	6 3	
Tacrolimus	Accuracy Completeness	5 3	5 2	
IVIG	Accuracy Completeness	5 2	6 2	
Total mean (SD)	Accuracy Completeness	5 (1.17) 2.5 (0.51)	5.07 (1.26) 2.64 (0.49)	*P1*=0.769¶ *P2*=0.541¶
Total scores of all items	Mean (SD) Median (IQR)	3.75 (1.55) 3 (4)	3.86 (1.57) 3 (4)	*P*=0.717¶

GPT: Generative Pre-trained Transformer; NSAID: non-steroidal anti-inflammatory drug; IVIG: intravenous immunoglobulin; SD: Standard deviation, IQR: Interquartile range; *P*1: *P*-value of accuracy scores, *P*2: *P*-value of completeness scores; *P*<0.05 was considered statistically significant. ¶: Mann–Whitney U test

The frequency of the accuracy and completeness scores of answers generated by two GPT versions were evaluated. For accuracy, GPT 4 yielded scores of 7.1% (n=1) more incorrect than correct, 28.6% (n=4) more correct than incorrect, 7.1% (n=1) nearly all correct, and 57.1% (n=8) correct. GPT 3.5 yielded scores of 7.1% (n=1) more incorrect than correct, 21.4% (n=3) more correct than incorrect, 28.6% (n=4) nearly all correct, and 42.9% (n=6) as shown in Figure 1. For completeness, GPT 4 yielded scores of 64.3% (n=9) comprehensive and 35.7% (n=5) adequate, and no incomplete score was noted. GPT 3.5 yielded scores of 50.0% (n=7) comprehensive and 50.0% (n=7) as adequate, and no incomplete score was noted (Figure 2).

Figure 1: Distribution of the accuracy scores for Chat Generative Pre-trained Transformer (ChatGPT) versions 3.5 and 4

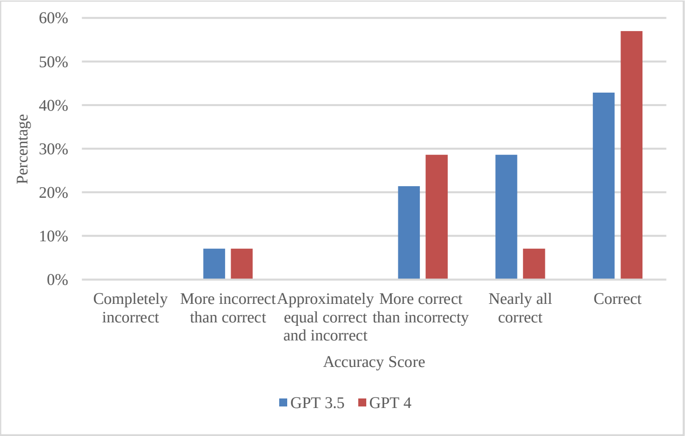
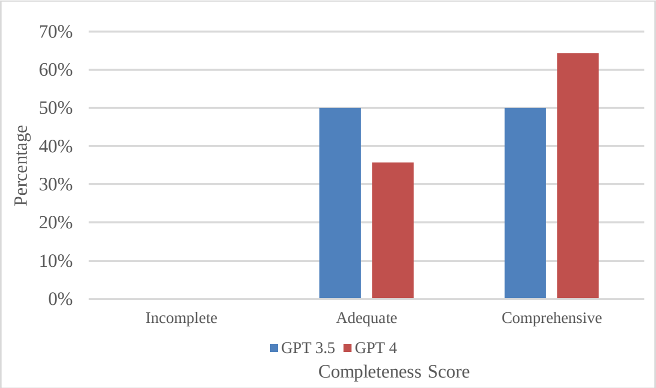


Figure 2: Distribution of the completeness scores for ChatGPT versions 3.5 and 4.



The inter-rater reliability as assessed by Cohen's kappa coefficient demonstrated a level of agreement ranging from good to very good. The agreement of the accuracy scores for GPT versions 3.5 and 4 were 0.794 (*P*<0.001) and 0.763 (*P*<0.001), respectively. The agreement of the completeness scores for GPT 3.5 and GPT 4 were 0.714 (*P*=0.008) and 0.851 (*P*=0.001), respectively.

In the GPT 3.5 model, hydroxychloroquine and Leflunomide received the highest full scores for both accuracy and completeness, while methotrexate, Sulfasalazine, Cyclophosphamide, Mycophenolate mofetil, and Tofacitinib received the highest total scores in the GPT 4 model. Azathioprine received the lowest accuracy and completeness scores for the GPT 3.5 model and Ciclosporin for the GPT 4 model.

Discussion

The popularity of AI use, particularly ChatGPT, in the field of healthcare is increasing. However, data concerning its reliability and adequacy are still not entirely sufficient. This study aimed to evaluate the accuracy and completeness of ChatGPT version 3.5 (free) and version 4 (fee for use) in the context of digital health-related information on the use of anti-rheumatic medicines in pregnancy. Based on the results of our research, the answers generated by ChatGPT versions 3.5 and 4 to inquiries about the safety and compatibility of rheumatological medications during pregnancy demonstrate a satisfactory level of accuracy and completeness. The outcomes of both versions exhibited similarities and did not demonstrate superiority over one another. To the best of our knowledge, our study is the first study in which ChatGPT versions 3.5 and 4 were evaluated in terms of the use of anti-rheumatic medicines in pregnancy, and no similar study was found in the literature. From the patient's perspective, pregnancy while using rheumatological medicines is a subject that involves a high level of anxiety and motivates patients to investigate this issue. Our study is important in terms of evaluating whether patients have access to accurate and sufficient information other than their physicians.

Since the potential teratogenic effect of many medicines has not yet been demonstrated, the use of medicines in pregnancy should be approached carefully [11]. It has been reported that only 5% of 213 new medicines approved by the United States Food and Drug Administration (USFDA) between 2003 and 2012 can be used safely in pregnancy, and information on whether many medicines can be used safely in pregnancy is still limited [12,13]. This situation causes anxiety in pregnant women who have chronic diseases, such as rheumatological diseases, and need to continue pharmacological therapy during pregnancy. Therefore, easily accessible sources of information come into play at this stage and are used to help patients find answers to their questions. Studies show that the rate of pregnant women’s use of Internet resources related to medicine use reaches as high as 76%, and the Internet plays an essential role in pregnant women's access to health information and decision-making [14,15].

To what extent can the information about medicine and health found on the internet be deemed accurate? Which application or website provides the most accurate and trustworthy data? The proliferation of the Internet in the healthcare industry has prompted these questions. ChatGPT is a popular and generally trustworthy model of artificial intelligence. Sabry Abdel-Messih et al. [16] investigated the capabilities of ChatGPT to respond to questions regarding a specific case of acute organophosphate poisoning in their research. That study’s findings demonstrated that the model effectively addressed all questions posed. Both the initial and reconstructed responses obtained from ChatGPT were deemed to be highly satisfactory. They stated that as ChatGPT evolves and its application in medicine becomes more refined, AI could be useful for addressing rare clinical cases, which are sometimes overlooked by experts, as opposed to replacing healthcare professionals. Similarly, it has been reported to be a useful tool in many medical areas, such as cirrhosis and hepatocellular carcinoma, dental applications, drafting, and plastic surgery [17–19].

In addition to the data supporting the reliability and adequacy of ChatGPT versions, studies claiming the opposite have also been published. In the study by Jeblick et al. [20] in which radiology reports were evaluated, potentially harmful errors, such as missing important medical findings, were identified, and they emphasized the need for manual checking of these automated reports. In the discharge summary example provided in a study by Patel [20], ChatGPT added extra information to the summary that was not included in the prompts [21]. In a study by Alkaissi et al., questions about homocysteine were asked to ChatGPT, and although they received mostly correct answers, they also received irrelevant answers. When asked to provide references on this subject, the ChatGPT provided article titles that did not exist. The PubMed IDs he provided for these articles were for completely different and unrelated articles [22]. So, how does ChatGPT provide information that does not exist? As far as we know, chatbots respond to pre-programmed datasets. However, generative models, such as ChatGPT, can generate new information that is not real. Alkaissi et al. [22] called this condition an “artificial hallucination”. This artificial condition raises concerns about the level of ChatGPT’s reliability.

One important finding of our study is that ChatGPT generally performed better with medications, such as methotrexate and Leflunomide, which are contraindicated in pregnancy and whose association with malformations is well-established. Although other drugs generally did not reach full accuracy and completeness scores, it was emphasized that decisions should be made according to the condition of the patient and his/her disease status in addition to the benefit/risk ratio obtained from the ChatGPT. Additionally, it was stated that healthcare professionals should be consulted before deciding whether (or not) to use such a tool.

Limitations

This study has some limitations. First, this study was designed to evaluate the existing versions of ChatGPT. The database used to train ChatGPT only contains information through 2021. Due to this limitation, the information that is provided in the study may not be current. The study is conducted solely in English, which may not fully represent the AI’s capabilities in other languages or the global diversity of users seeking information on anti-rheumatic drugs during pregnancy. The subjective nature of Likert-type scales and self-reported scores may result in bias. Additionally, previous experiences or preconceived notions of the investigators regarding the use of anti-rheumatic drugs during pregnancy may influence their evaluation and lead to bias.

Further research is needed to better investigate ChatGPT’s reliability and comprehensiveness across different medical fields. Additionally, more comprehensive studies should be done to evaluate whether this tool produces the same results in other languages.

Conclusion

In conclusion, while ChatGPT versions 3.5 and 4 offer a substantial amount of reliable information, the prevailing research indicates the necessity of acknowledging the limitations inherent in the information derived from these models. This study demonstrated that the AI chatbots, GPT versions 3.5 and 4 provide accurate and comprehensive information to patients in the setting of anti-rheumatic drug use during pregnancy. On the other hand, information generated by ChatGPT must be validated, and patients should be cautioned about the potential of receiving misinformation on health-related issues. Evaluation and advancement of these tools are essential steps for assuring the accuracy and quality of the information generated. Due to ChatGPT’s limitations, it cannot serve as a stand-alone decision-making tool for such a sensitive issue, such as the use of medication during pregnancy. ChatGPT lacks access to and cannot analyze a patient’s laboratory parameters, prior pregnancy complications, and the internal dynamics of the patient. The information acquired from ChatGPT necessitates verification. While both iterations of ChatGPT offer valuable insights, it is crucial to keep in mind that ChatGPT does not possess the expertise of a medical professional. Further research is required to investigate and develop its potential for use in treating a variety of medical conditions.

References

1. Cooper GS, Stroehla BC. The epidemiology of autoimmune diseases. Autoimmun Rev. 2003;2(3):119-25. doi: 10.1016/s1568-9972(03)00006-5.
2. Desai RJ, Huybrechts KF, Bateman BT, Hernandez-Diaz S, Mogun H, Gopalakrishnan C, et al. Brief Report: Patterns and Secular Trends in Use of Immunomodulatory Agents During Pregnancy in Women With Rheumatic Conditions. Arthritis Rheumatol. 2016;68(5):1183-9. doi: 10.1002/art.39521.
3. Grimes HA, Forster DA, Newton MS. Sources of information used by women during pregnancy to meet their information needs. Midwifery. 2014;30(1):e26-33. doi: 10.1016/j.midw.2013.10.007.
4. Serçekuş P, Değirmenciler B, Özkan S. Internet use by pregnant women seeking childbirth information. J Gynecol Obstet Hum Reprod. 2021;50(8):102144. doi: 10.1016/j.jogoh.2021.102144.
5. Bramham K, Soh MC, Nelson-Piercy C. Pregnancy and renal outcomes in lupus nephritis: an update and guide to management. Lupus. 2012;21(12):1271-83. doi: 10.1177/0961203312456893.
6. Pal S, Bhattacharya M, Lee SS, Chakraborty C. A Domain-Specific Next-Generation Large Language Model (LLM) or ChatGPT is Required for Biomedical Engineering and Research. Ann Biomed Eng. 2023;10. doi: 10.1007/s10439-023-03306-x.
7. Deiana G, Dettori M, Arghittu A, Azara A, Gabutti G, Castiglia P. Artificial Intelligence and Public Health: Evaluating ChatGPT Responses to Vaccination Myths and Misconceptions. Vaccines (Basel). 2023 7;11(7):1217. doi: 10.3390/vaccines11071217.

8. Deng J, Lin Y. The Benefits and Challenges of ChatGPT: An Overview. Frontiers in Computing and Intelligent Systems. 2023;2(2) 81-83. doi: 10.54097/fcis.v2i2.4465.
9. Götestam SC, Hoeltzenbein M, Tincani A, Fischer-Betz R, Elefant E, Chambers C, et al. The EULAR points to consider for use of antirheumatic drugs before pregnancy, and during pregnancy and lactation. Ann Rheum Dis. 2016;75(5):795-810. doi: 10.1136/annrheumdis-2015-208840.
10. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Research square. 2023;28:rs.3.rs-2566942. doi: 10.21203/rs.3.rs-2566942/v1.
11. Olukman M, Parlar A, Orhan CE, Erol A. Gebelerde ilaç kullanımı: Son bir yıllık deneyim. Turkish Journal of Obstetrics and Gynecology. 2006;3(4):255-61. doi:10.17049/ataunihem.499684.
12. Riley LE, Cahill AG, Beigi R, Savich R, Scade G. Improving safe and effective use of medicines in pregnancy and lactation. American Journal of Perinatology. 2017;34(8):826-32. doi: 10.1055/s-0037-1598070.
13. Oliveire-Filho A, Veire AES, Silvo RC, Neves STF, Gama TAB, Lima RV, et al. Adverse medicine reactions in high-risk pregnant women. Saudi Pharmaceutical Journal. 2017;25(7):1073-7. doi: 10.1016/j.jsps.2017.01.005.
14. Sinclair M, Lagan BM, Dolk H, McCullough J. An assessment of pregnant women's knowledge and use of the internet for medication safety information and purchase. Journal of Advanced Nursing. 2018;74(1):137-47. doi: 10.1111/jan.13387.
15. Koyun A, Kesim Sİ. Gebelikte Karar Vermeye İnternetin Etkisi: Sistematik Bir İnceleme. 3. Uluslararası Bilimsel Araştırmalar Kongresi Bildiri Kitabı, 2018: 9-23.
16. Sabry Abdel-Messih M, Kamel Boulos MN. ChatGPT in Clinical Toxicology. JMIR Med. Educ. 2023;9:e46876. doi: 10.2196/46876.
17. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol. 2023;29(3):721-32. doi: 10.3350/cmh.2023.0089.
18. Alhaidry HM, Fatani B, Alrayes JO, Almana AM, Alfhaed NK. ChatGPT in Dentistry: A Comprehensive Review. Cureus. 2023;15(4):e38317. doi: 10.7759/cureus.38317.
19. Sharma SC, Ramchandani JP, Thakker A, Lahiri A. ChatGPT in Plastic and Reconstructive Surgery. Indian J Plast Surg. 2023;56(4):320-5. doi: 10.1055/s-0043-1771514.
20. Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, et al. ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports. 2022;10.48550/arXiv.2212.14882.
21. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health. 2023;5(3):e107-8. doi: 10.1016/S2589-7500(23)00021-3.
22. Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. Cureus. 2023;15(2):e35179. doi: 10.7759/cureus.35179.