Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–7

DOI: 10.1177/20552076231224603
journals.sagepub.com/home/dhj

Sage

Evaluation of information provided to patients by ChatGPT about chronic diseases in Spanish language

María Juliana Soto-Chávez[1], Marlon Mauricio Bustos[1,2], Daniel G. Fernández-Ávila[1,3] and Oscar Mauricio Muñoz[1,2]

Abstract

Introduction: Artificial intelligence has presented exponential growth in medicine. The ChatGPT language model has been highlighted as a possible source of patient information. This study evaluates the reliability and readability of ChatGPT-generated patient information on chronic diseases in Spanish.

Methods: Questions frequently asked by patients on the internet about diabetes mellitus, heart failure, rheumatoid arthritis (RA), chronic kidney disease (CKD), and systemic lupus erythematosus (SLE) were submitted to ChatGPT. Reliability was assessed by rating responses as (1) comprehensive, (2) correct but inadequate, (3) some correct and some incorrect, (4) completely incorrect, and divided between "good" (1 and 2) and "bad" (3 and 4). Readability was evaluated with the adapted Flesch and Szigriszt formulas.

Results: And 71.67% of the answers were "good," with none qualified as "completely incorrect." Better reliability was observed in questions on diabetes and RA versus heart failure ($p = 0.02$). In readability, responses were "moderately difficult" (54.73, interquartile range (IQR) 51.59–58.58), with better results for CKD (median 56.1, IQR 53.5–59.1) and RA (56.4, IQR 53.7–60.7), than for heart failure responses (median 50.6, IQR 46.3–53.8).

Conclusion: Our study suggests that the ChatGPT tool can be a reliable source of information in spanish for patients with chronic diseases with different reliability for some of them, however, it needs to improve the readability of its answers to be recommended as a useful tool for patients.

Keywords

Artificial intelligence, ChatGPT, chronic diseases, reliability, readability

Submission date: 19 July 2023; Acceptance date: 18 December 2023

Introduction

Artificial intelligence (AI) has shown a remarkable growth since its conceptualization in 1950, being an emerging technology with applications in different fields such as medicine.[1,2] In the last decade, the development of machine learning has given way to language models such as Open AI's recently launched pretrained generative chatbot: ChatGPT. AI-based language models are promising tools in medicine given their wide availability, ability to integrate information and automate activities, which could improve the screening, diagnosis, and treatment of different diseases.[3]

Currently, patients are often looking for information about their diseases on the Internet through search engines such as Google. It is known that up to 75% of patients with

[1]Department of Internal Medicine, Pontificia Universidad Javeriana, Bogotá, Colombia
[2]Department of Internal Medicine, Hospital Universitario San Ignacio, Bogotá, Colombia
[3]Rheumatology Unit, Hospital Universitario San Ignacio, Bogotá, Colombia

Corresponding author:
María Juliana Soto-Chávez, Department of Internal Medicine, Hospital San Ignacio, Carrera 7 40–62, Bogotá, Colombia.
Email: msotoc@javeriana.edu.co

chronic diseases are taking decisions related to their pathologies based on information found on the Internet.[4] However, recent studies have shown that the use of tools such as ChatGPT has been associated with positive user experiences in obtaining answers about their pathologies,[5] so a rapid growth in its use is expected.

Experts in hepatopathies and bariatric surgery have evaluated the reliability and reproducibility of information generated by ChatGPT in English, highlighting its usefulness for patients.[6,7] However, to our knowledge, there are no studies that have evaluated the information provided by language models such as ChatGPT in chronic diseases or any other specific conditions to provide medical information to patients in Spanish.

The aim of this study is to evaluate the reliability and readability in Spanish of the information presented by ChatGPT for patients on different chronic diseases such as diabetes mellitus, heart failure, rheumatoid arthritis (RA), chronic kidney disease (CKD), and systemic lupus erythematosus (SLE).

Methods

Data collection

An analytical cross-sectional observational study was conducted, with the aim of evaluating the reliability and readability in Spanish of the information provided by ChatGPT. The STROBE checklist for cross-sectional studies was used to guide the report. No informed consent was required given the nature of the study. Questions frequently asked by patients on health forums, social media (Facebook, Twitter/X), and search engines such as Google about general data, diagnosis, nonpharmacological and pharmacological treatment, and complications, were included. Questions with ambiguity, unverifiable information, no clear or objective answer were excluded. Twelve questions standardized by the researchers were selected to be applied to the five chronic diseases included in the study (Supplemental Table 1). The questions were classified into five categories: (1) general information about the disease; (2) diagnosis; (3) nonpharmacological treatment; (4) pharmacological treatment; and (5) complications (Supplemental Table 1). The study was considered an investigation without risk and was approved by the ethics committee of the Pontificia Universidad Javeriana (FM-CIE-0533-23).

ChatGPT response generation

ChatGPT is a natural language processing model, released for public use in November 2022. With data obtained from web sources, books, and scientific articles until 2021, the model generates answers in a conversational manner, incorporating user feedback, and correction.[5,6] The ChatGPT's Free Research Preview (May 12 version) was used to generate the questions. A user was created in "incognito mode" for the present study. Each question was asked separately and independently using the "New Chat" function. The first response generated by the chatbot was considered for grading, with no response regeneration.

Response evaluation

The generated responses were independently evaluated by two internal medicine physicians (MJS and MMB). The reliability of the answers was evaluated by a scale used in previous studies to grade the information generated by ChatGPT[6,7]: (1) comprehensive (similar to an answer that a specialist in the subject would give); (2) correct but inadequate (accurate but incomplete information, missing some important information); (3) some correct and some incorrect; and (4) Completely incorrect.[7,8] An evaluation of the interobserver correlation of the answers was carried out with Cohen's kappa coefficient, determining a fair agreement (0.2241; CI 0.13–0.32), and therefore a consensual qualification was generated among the investigators in the ratings with differences.

Finally, the readability of the answers generated was evaluated by means of the Flesch formula adapted to Spanish and the Flesch-Szigriszt readability formula,[9,10] which evaluate the legibility of health information for patients.

Statistical analysis

We computed the sample size required to detect discordant proportions of 0.1 and 0.3 (delta 0.2) for a two-sample paired-proportions test, with 80% power using a two-sided test and 5% significance level. The calculated sample size was 60 questions, 12 for each disease.

Proportions of reliability ratings were determined for each answer category and for each disease. To synthesize the information, the ratings were dichotomized as "good" ("comprehensive" or "correct but inadequate") or "bad" ("some correct and some incorrect" or "completely incorrect"). For readability, median and interquartile ranges are presented since the distribution of their data was not normal according to the Shapiro-Wilk test.

Significant differences between the "good" answers by disease and by answer category were determined using the chi-square test. The Kruskal-Wallis test was used for comparison of readability between the different diseases and categories of answers. Values of $p < 0.05$ were considered statistically significant. The statistical program Stata (16.1, StataCorp LLC, College Station, TX, USA) was used for the analysis.

Results

Reliability

Of the 60 generated responses by ChatGPT, 41.6% were top rated as "comprehensive," and 71.67% were considered "good." Within the answers analyzed, there were none rated as "completely incorrect." Regarding the evaluation by

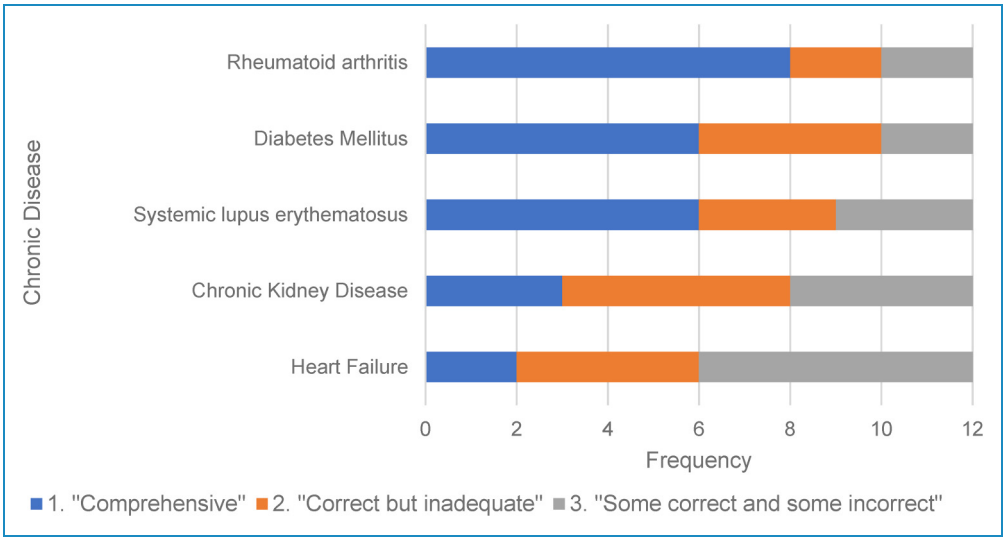

Figure 1. Qualification of responses generated by ChatGPT about chronic diseases.

Table 1. Reliability rating of ChatGPT-generated responses rated as "good" ("comprehensive" or "correct but inadequate").

	Chronic disease					
	Diabetes, n = 12	Heart failure, n = 12	CKD, n = 12	Rheumatoid arthritis, n = 12	SLE, n = 12	Total by category, n (%)[b]
Category						
General information, n = 4[a] (%)	3 (75)	3 (75)	3 (75)	3 (75)	3 (75)	15 (75)
Diagnosis, n = 2[a] (%)	1 (50)	0 (0%)	1 (50)	2 (100)	2 (100)	6 (60)
Nonpharmacological treatment, n = 2[a] (%)	2 (100)	1 (50)	2 (100)	1 (50)	1 (50)	7 (70)
Pharmacological treatment, n = 3[a] (%)	3 (100)	2 (66.7)	2 (66.7)	3 (100)	2 (66.7)	12 (80)
Complications, n = 1[a] (%)	1 (100)	0 (0)	0 (0)	1 (100)	1 (100)	3 (60)
Total by disease, n (%)	10 (83.3)	6 (50)	8 (66.6)	10 (83.3)	9 (75)	

aResponses in the category by disease.
bPercentage in relation to the total questions in the category.
CKD = chronic kidney disease; SLE = systemic lupus erythematosus.

disease, RA obtained the highest ratings, while CKD and heart failure received the lowest ratings, as shown in Figure 1. Furthermore, a statistically significant difference was observed in the performance of ChatGPT in generating answers considered "good" for questions related to diabetes and rheumatoid arthritis, compared to heart failure questions (83% vs. 50%, p = 0.02), as shown in Table 1.

In terms of question categories, ChatGPT showed superior performance in the pharmacological treatment category, with 80% of answers rated as "good." On the other hand, inferior performance was evidenced in the diagnosis and complications categories, with 60% of answers rated as "good."

Readability

The readability level on the information provided by ChatGPT to on chronic diseases was evaluated. Using

Table 2. Fleshner readability rating of chronic disease ChatGPT responses.

	Chronic disease median (IQR)					
	Diabetes, n = 12	Heart failure, n = 12	CKD, n = 12	Rheumatoid arthritis, n = 12	SLE, n = 12	Total by category, median (IQR)
Category, median (IQR)						
General information	56.5 (54.5–58.1)	51.1 (43.1–53.3)	57.0 (54.3–63.8)	60.8 (50.5–62.4)	56.0 (53.1–59.9)	55.2 (53.1–59.4)
Diagnosis	68.3 (60.5–76.1)	54.5(50.8–58.2)	62.8 (58.6–67.0)	60.7 (60.2–61.2)	60.4 (55.7–65.2)	60.4 (58.2–65–2)
Nonpharmacological treatment	44.4 (34.2–54.1)	45.5(37.5–53.5)	53.7(51.1–56.4)	54.7(54.4–55.1)	51.3(47.6–55.0)	53.8 (47.6–55.0)
Pharmacological treatment	55.6(53.6–55.7)	47.9(44.7–54.5)	55.8(50.9–59.6)	54.0(53.4–57.6)	51.4(47.5–58.7)	54.0 (50.9–55.8)
Complications	56.2	50.4	53.8	51.6	51.3	51.6 (51.3–53.9)
Total by disease	55.6 (54.1–58.1)	50.6 (46.3–53.8)	56.1 (53.5–59.1)	56.4 (53.7–60.7)	54.1 (51.4–58.8)	

CKD = chronic kidney disease; IQR = interquartile range; SLE = systemic lupus erythematosus.

Fleshner's formula adapted to Spanish, a median readability of 54.73 (interquartile range (IQR) 51.59–58.58) was obtained for the total set of answers generated by ChatGPT, being classified as "moderately difficult."

When evaluated by disease, better readability was obtained for the answers on CKD (median 56.1, IQR 53.5–59.1) and RA (56.4, IQR 53.7–60.7), than for the answers on heart failure (median 50.6, IQR 46.3–53.8) as shown in Table 2, a difference that was statistically significant ($p = 0.016$).

Evaluation by categories showed a higher readability for diagnosis than for complications (median 60.4, IQR 58.2–65–2 vs 51.6, IQR 51.3–53.9, $p = 0.005$).

When assessing the readability with the Szigriszt perspicuity formula, similar results were shown when assessed by diseases, with a better readability for CKD and RA when compared to heart failure ($p = 0.022$); and by categories, showing better readability for diagnosis and lower for complications ($p = 0.006$) (Supplemental Table 2).

Discussion

To our knowledge, this is the first study to evaluate the information generated by an AI-based language model for patients, in Spanish language. In addition, it is the first study to compare the generated responses by ChatGPT for different chronic diseases. Our study suggests that most of the answers generated responses by ChatGPT are considered "good," with none qualified as "completely incorrect," Additionally, we found differences among the chronic diseases evaluated, being better for RA. Notably, the generated responses are generally not properly readable.

Search engines such as Google and social networks such as YouTube and Facebook have become widely used resources by patients in search of information about their health condition, with high variability in the quality of information found through these media,[11–14] and often with a low readability.[15,16] Given the possibility that AI-based language processing models may improve the accessibility of information for patients, we decided to evaluate the information provided by ChatGPT on chronic diseases.

The responses generated by ChatGPT were assessed by two investigators who rated them independently. A weak concordance between the ratings was found, indicating that healthcare professionals may differ on what information should be conveyed to patients. This highlights the importance of providing recommendations not only regarding clinical practice, but also the way information should be presented to patients. We emphasize that the presentation of information may vary depending on the audience to which is presented.

In our study, the majority of answers were rated as "good," similar to findings in studies that evaluated the information provided on liver disease and bariatric surgery in English with 79.1% and 86.8%, respectively.[6,7] Given the importance and benefits of health literacy, our results suggest that AI-based languages are resources that would allow the patients to obtain accurate and personalized information about their condition. ChatGPT can therefore provide a tool to improve access to potentially reliable and accurate information about chronic conditions.

Responses generated on diabetes and RA were better qualified especially in terms of pharmacological and non-pharmacological treatment (100% of "good" answers).

Although these diseases differ in terms of prevalence, diagnosis and treatment, the answers were more accurate, better explained and based on the best available evidence. The answers with the lowest qualification were on heart failure, especially in diagnosis and complications categories (0% of "good" answers). We noted that these answers were more generic, being unspecific and outdated even for the year 2021, the date by which the data used by ChatGPT has been updated,[6] which may lead to a delay in considering recent and innovative diagnostic alternatives. We also found that the reliability is lower for information provided on more prevalent chronic diseases such as heart failure and CKD, that could be explained by the large amount of information available in its databases about these diseases. We hypothesize that such a high volume of information may hinder the process of selecting and filtering the information that will finally be delivered to the patient.

Regarding the categories of answers, the chatbot presents a better performance in the category of pharmacological treatment (80% of "good" answers), being inferior in the categories of diagnosis and complications (60% of "good" answers). Such differences can be explained by the fact that the tool occasionally presents outdated and currently unaccepted information, with incomplete answers and even incorrect data in the latter categories. It is worth noting that it frequently recommends the user to seek advice from an expert in each disease, to adequately determine its diagnosis and individualized treatment.

On the other hand, we evaluated readability with two available formulas that have been validated in Spanish. In general, the tool provides answers with a readability rated as "moderately difficult," which differs from the recommendations for readability of medical information in Spanish,[17] something that should be considered a limitation for recommending it as a source of information for patients with chronic diseases.

The best readability was obtained in CKD and RA, possibly because these diseases use less complex and have a more precise medical terminology. In addition, we found significant differences in readability by category, being higher for diagnosis and lower for complications, maybe because for the diagnosis of chronic diseases the tool uses well-defined and simpler concepts compared to those used for complications.

It is important to consider that there are differences in the performance of ChatGPT depending on the language in which the question is asked. Models as ChatGPT are capable of processing and generating text in multiple languages to some extent, although their performance tends to be better in the languages they were trained on. Since this tool has been trained in English and is mostly based on databases available in this language, the answers in Spanish could be less accurate, as clarified by its creators.[6] A preliminary evaluation of ChatGPT for machine translation in business topics, including translation speed, multilingual translation, and translation robustness, by Wenxiang et al., showed that ChatGPT competes with commercial translation products (e.g. Google Translate) in high-resource European languages, but lags far behind in distant or low-resource languages.[18] Although to date there are studies that have evaluated the performance of the ChatGPT tool in national medical validation exams in Japan, Brazil, and Spain in their respective languages,[19–21] there have been no studies comparing the accuracy of the answers in English versus the answers in Spanish in terms of medical information generated by the tool. Additionally, we believe the use of the Roman alphabet and European languages might yield different results due to variations in linguistic structures and medical practices. This area could be the subject of future research.

Another aspect that can influence the generated responses is the way the question is formulated. It has been shown that when typical consultation questions are asked in an objective manner (e.g. how is the diagnosis of hypertension made?), the information presented by the chatbot usually provides useful answers that help both the patient and the health professional; however, when questions are asked in a subjective or non-specific manner (e.g. What do I take for high blood sugar?), less individualized and accurate answers are obtained,[8,22] which should be taken into account when evaluating the recommendation of this tool for patient education. Taking into account that the standardized questions chosen by the investigators were direct and carefully formulated, we consider that our findings probably apply to a population with higher academic levels and with some basic knowledge about their disease. This limitation is similar to the reported for patient education materials developed by multiple medical associations where the readability is suboptimal, as showed by Minh who found that those materials are written at a level above the recommended sixth grade reading level recommended by the Centers for Disease Control and National Institutes of Health.[17] Future research is needed to evaluate the impact of the phrasing of questions and the educational level in the reliability and readability of information.

Limitations and strengths

In the present study we highlight several strengths. First, it includes questions and answers in Spanish, which has not been previously evaluated. Second, we compared reliability and readability, which provides a more complete view of the performance, quality, and clarity of the tool for presenting medical information on chronic diseases. Third, the different questions were classified by categories, determining that there are no significant differences between them. Also, although a larger number of raters could offer a more comprehensive evaluation of the answers, our study has already achieved a fair agreement between evaluators as mentioned previously, which we believe demonstrate a solid level of consensus in the assessments.

There are some limitations that we should mention. First, we must recognize that ChatGPT is a tool in constant evolution, so the answers may vary depending on the context, the version of the AI, and the time at which the question is

asked. Likewise, we should note that we only evaluated the first answer generated by the chatbot, without considering the "regenerated responses," which could have omitted information that was corrected or supplemented in subsequent answers. These aspects should be considered when interpreting the results of our research.

It should also be noted as a limitation in our study that the interpretation of the answers provided by ChatGPT is made from the point of view of health personnel (physicians), which could differ from the perception of patients, where specific characteristics such as age and level of education should be considered, as they could affect the degree of understanding of the answers provided, as well as the capacity and ability to use the tool. We must also consider the differences in patients' levels of health literacy, which can vary between different populations and locations. These characteristics should be evaluated in future research to increase knowledge of the reliability of AI tools.

Finally, our study did not compare traditional search engines to the use of ChatGPT. Those engines provide a very extensive information, which becomes overwhelming to patients, and makes it difficult to evaluate the reliability and readability of content from those sources. Future research should be conducted to compare factors such as accuracy, relevance, and usability of information retrieval using ChatGPT and traditional search engines.

Conclusions

Our study suggests that ChatGPT-generated responses on chronic diseases have good reliability, with none qualified as "completely incorrect." However, there are differences in performance according to disease type. The answers on diabetes and RA evidenced better performance, with a lower score on the heart failure answers. In terms of readability, we found that the responses were rated as "moderately difficult," with heart failure answers being the most difficult to read. Although ChatGPT can be a reliable source of information in Spanish for patients with chronic diseases, the readability of its answers should be improved to recommend it as a useful patient education tool.

Contributorship: MS, OM, and DF were involved in study conception and design; MS and MB in data collection; MS, OM, and MB in analysis and interpretation of results; MS in draft manuscript preparation; and OM, MB, and DF in editing and completion of the manuscript. All authors reviewed the results and approved the final version of the manuscript.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: The study protocol was approved by the Hospital Universitario San Ignacio Ethics Committee.

Funding: The authors received no financial support for the research, authorship, and/or publication of this article.

Gurantor: MJS

ORCID iD: María Juliana Soto-Chávez https://orcid.org/0000-0003-4946-8774
Oscar Mauricio Muñoz https://orcid.org/0000-0001-5401-0018

Supplemental material: Supplemental material for this article is available online.

References

1. IBM. What is artificial intelligence? [Internet]. 2023. Available from: https://www.ibm.com/topics/artificial-intelligence
2. Zaar O, Larson A, Polesie S, et al. Evaluation of the diagnostic accuracy of an online artificial intelligence application for skin disease diagnosis. *Acta Derm Venereol* 2020; 100: 1–6.
3. Davenport T and Kalakota R. The potential for artificial intelligence in healthcare. *Futur Healthc J* 2019; 6: 94–98.
4. Fox S. Online Health Search 2006 [Internet]. 2006 [cited 2022 Feb 26]. Available from: https://www.pewresearch.org/internet/2006/10/29/online-health-search-2006/
5. Salah M. Chatting with ChatGPT: decoding the mind of Chatbot users and unveiling the intricate connections between user perception, trust and stereotype perception on self-esteem and psychological well-being. *Curr Psychol* 2023: 1–26. Preprint. https://doi.org/10.21203/rs.3.rs-2610655/v2
6. AI O. Introducing ChatGPT. 2022; Available from: https://openai.com/blog/chatgpt
7. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol* 2023 Mar 22: 721–732. Available from: http://www.e-cmh.org/journal/view.php?doi=10.3350/cmh.2023.0089.
8. Samaan JS, Yeo YH, Rajeev N, et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. *Obes Surg* 2023: 1790–1796.
9. Rios Hernandez IN. Un acercamiento a la legibilidad de textos relacionados con el campo de la salud. *Chasqui Rev Latinoam Comun* 2017: 253.
10. Lorza DPQ, Morales VT and Méndez NDD. Herramienta para análisis de la legibilidad lingüística de contenido web en español como apoyo a procesos de evaluación de accesibilidad. *Brazilian Creat Ind J* 2021 Jul 1; 1: 71–87. Available from: https://periodicos.feevale.br/seer/index.php/braziliancreativeindustries/article/view/2677.
11. Keselman A, Arnott Smith C, Murcko AC, et al. Evaluating the quality of health information in a changing digital ecosystem. *J Med Internet Res* 2019 Feb; 21: e11129.
12. Barahona-correa E, Romero-alvernia DM and Rueda-ortiz C. Social media as source of information for Spanish-speaking patients with systemic lupus erythematosus. 2022; 31: 953–962.
13. Camm CF, Russell E, Ji-Xu A, et al. Does YouTube provide high-quality resources for patient education on atrial fibrillation ablation? *Int J Cardiol* 2018 Dec; 272: 189–193.

Available from: https://linkinghub.elsevier.com/retrieve/pii/S0167527318333382.
14. Mukewar S, Mani P, Wu X, et al. YouTube ®and inflammatory bowel disease. *J Crohn's Colitis*. 2013;7:392–402. Available from:
15. Oliffe M, Thompson E, Johnston J, et al. Assessing the readability and patient comprehension of rheumatology medicine information sheets: A cross-sectional Health Literacy Study. *BMJ Open* 2019 Feb 5; 9: e024582. Available from: https://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2018-024582.
16. Hoang PM and van Ballegooie C. Assessment of the readability and quality of online patient education material for chronic medical conditions. *Healthcare* 2022 Jan 26; 10: 234. Available from: https://www.mdpi.com/2227-9032/10/2/234.
17. Kloosterboer A, Yannuzzi NA, Patel NA, et al. Assessment of the quality, content, and readability of freely available online information for patients regarding diabetic retinopathy. *JAMA Ophthalmol* 2019 Nov 1; 137: 1240. Available from: https://jamanetwork.com/journals/jamaophthalmology/fullarticle/2748609.
18. Jiao W, Wang W, Huang J-T, et al. *Is ChatGPT a good translator? Yes with GPT-4 as the engine*. 2023. ArXiv Preprint. Available from: https://arxiv.org/abs/2301.08745
19. Takagi S, Watari T, Erabi A, et al. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. *JMIR Med Educ* 2023 Jun 29; 9: e48002. Available from: https://mededu.jmir.org/2023/1/e48002.
20. Gobira M, Nakayama LF, Moreira R, et al. Performance of ChatGPT-4 in answering questions from the Brazilian national examination for medical degree revalidation. *Rev Assoc Med Bras*. 2023;69: 1460–1487. Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-42302023001000618&tlng=en
21. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, et al. Evaluating the efficacy of ChatGPT in navigating the Spanish medical residency entrance examination (MIR): promising horizons for AI in clinical medicine. *Clin Pract* 2023 Nov 20; 13: 1460–1487. Available from: https://www.mdpi.com/2039-7283/13/6/130.
22. Lee P, Bubeck S and Petro J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. Drazen JM, Kohane IS, Leong T-Y, editors. *N Engl J Med* 2023 Mar 30; 388: 1233–1239. Available from: http://www.nejm.org/doi/10.1056/NEJMsr2214184.