Assessing the Readability, Quality and Reliability of Responses Produced by ChatGPT, Gemini, and Perplexity Regarding Most Frequently Asked Keywords about Low Back Pain

Erkan Ozduran[1], Volkan Hancı[2], Yüksel Erkin[3], İlhan Celil Özbek[4], Vugar Abdulkerimov[5]

[1]Sivas Numune Hospital, Physical Medicine and Rehabilitation, Pain Medicine, Sivas, Turkey
[2]Dokuz Eylul University, Anesthesiology and Reanimation, Critical Care Medicine, Izmir Turkey
[3]Dokuz Eylul University, Anesthesiology and Reanimation, Pain Medicine, Izmir, Turkey
[4]Health Science University, Derince Education and Research Hospital, Physical Medicine and Rehabilitation, Kocaeli, Turkey
[5]Central Clinical Hospital, Anesthesiology and Reanimation, Baku, Azerbaijan

Corresponding author: Erkan Ozduran, MD
erkanozduran@gmail.com
Sivas Numune Hospital, Department of Physical Medicine and Rehabilitation/Pain Medicine, Yeşilyurt, 58380 Sivas Merkez/Sivas, Türkiye Turkey
Phone: +90 346 215 08 44 Mobile Phone: 05557004029
ORCID: https://orcid.org/0000-0003-3425-313X

ABSTRACT

Background. Patients who are informed about the causes, pathophysiology, treatment and prevention of a disease are better able to participate in treatment procedures in the event of illness. Artificial intelligence (AI), which has gained popularity in recent years, is defined as the study of algorithms that provide machines with the ability to reason and perform cognitive functions, including object and word recognition, problem solving and decision making. This study aimed to examine the readability, reliability and quality of responses to frequently asked keywords about (Low Back Pain) LBP given by 3 different AI -based chatbots (ChatGPT, Perplexity and Gemini), which are popular applications in online information presentation today.

Methods. All 3 AI chatbots were asked the 25 most frequently used keywords related to LBP determined with the help of Google Trend. In order to prevent possible bias that could be created by the sequential processing of keywords in the answers given by the chatbots, the study was designed by providing input from different users (EO, VH) for each keyword. The readability of the responses given was determined with Simple Measure of Gobbledygook (SMOG), Flesch Reading Ease Score (FRES) and Gunning Fog (GFG) readability scores. Quality was assessed using Global Quality Score (GQS) and Ensuring Quality Information for Patients (EQIP) score. Reliability was assessed by determining with DISCERN and Journal of American Medical Association (JAMA) scales.

Results. The first 3 keywords detected as a result of Google Trend search were "Lower Back Pain", "ICD 10 Low Back Pain", and "Low Back Pain Symptoms". It was determined that the readability of the responses given by all AI chatbots was higher than the recommended 6th grade readability level ($p < 0.001$). In the EQIP, JAMA, modified DISCERN and GQS score evaluation, Perplexity was found to have significantly higher scores than other chatbots ($P < 0.001$).

Conclusion. It has been determined that the answers given by AI chatbots to keywords about LBP are difficult to read and have low reliability and quality assessment. It is clear that when new chatbots are introduced, they can provide better guidance to patients with increased clarity and text quality. This study can provide inspiration for future studies on improving the algorithms and responses of AI chatbots.

Keywords: Artificial intelligence, ChatGPT, Gemini, Low back pain, Online medical information, Perplexity

Commented [JA1]: Why low back pain between brackets

INTRODUCTION

Low back pain (LBP) is a very common symptom and affects people of almost every age group. It is stated that the point prevalence of LBP that limits activity is 7.3% and 540 million people suffer from this complaint at some point in their lives. Not only that, it is emphasized that LBP is the number one cause of disability globally (Hartvigsen et al., 2018). In a study conducted by the Journal of the American Medical Association, it was determined that the expenditure on spine- related problems is the most costly expenditure after diabetes and heart disease. While medications, invasive procedures, imaging, and surgeries constitute direct related costs, disability, loss of productivity, and loss of wages are stated as indirect costs (Hemmer, 2021). The causes of LBP can often be mechanical, as well as chronic inflammatory diseases such as Ankylosing Spondylitis, which affects a rate of 0.1%-1.4% of the population (Bagcier,Yurdakul & Ozduran, 2021). According to the Centers for Disease Control and Prevention in the USA, in 2016, there were 3.6 million visits to emergency departments and 5.7 million visits to urgent and ambulatory care clinics due to back-related complaints (DePalma, 2020). Infection, fracture or trauma, malignancy, etc. are conditions that suggest urgent pathology. These conditions are called red flags, and failure to diagnose this condition by clinicians can lead to delayed treatment and increased patient morbidity and mortality (Verhagen et al., 2016).

Artificial intelligence (AI) can be defined as the study of algorithms that provide machines with the ability to reason and perform cognitive functions, including object and word recognition, problem solving, and decision making (Grippaudo et al., 2024). Artificial intelligence has gained popularity in recent years. Studies in the literature emphasize that the use of artificial intelligence robots that enable people to interact with technology in a more social and conversational manner is increasing. (Grippaudo et al., 2024; Gül et al., 2024) An example of conversational artificial intelligence is ChatGPT, developed by OpenAI (San Francisco, USA). It is widely used in many fields, especially in medical fields, and its reliability and effectiveness have been evaluated in many studies (Gül et al., 2024; Şahin et al., 2024). Perplexity AI is an artificial intelligence model that provides answers to queries and directions and includes links to quotations and related topics, while Google Gemini is an artificial intelligence model capable of analyzing complex data sets such as images and graphs (Ömür Arça et al., 2024).

Patients diagnosed with chronic nonspecific LBP do not have sufficient information about the amount and type of physical activity they can perform for their treatment. Therefore, patient

Commented [JA2]: Reference should be after the reference

education and back school programs for patients with LBP can help with spine protection, rehabilitation and the acquisition of specific information about the disease. (Ács et al., 2020; Nolet et al., 2018). As it is known, individuals who have better disease-specific knowledge, accurate information about the cause of the disease, prevention and treatment options have higher rates of protection from the disease and their participation in rehabilitation programs has also been determined to be higher (Járomi et al., 2021). In addition, the acquisition of health information via the internet is increasing day by day. In particular, people with LBP constantly express their desire to receive reliable information about their clinical condition (Hodges, Setchell & Nielsen, 2020). Patients can use popular internet search engines as well as artificial intelligence chatbots to obtain information in this area (Yilmaz Muluk & Olcucu, 2024). Artificial intelligence can also be used to monitor and give recommendations to patients experiencing chronic back pain. It can be used as an application that can be installed on mobile devices to monitor patients' symptoms and activities (Do et al., 2023). Hartmann et al., (2023) found a significant decrease in pain and pain-related impairments in daily living in patients diagnosed with LBP who used the AI-supported exercise application for 8 weeks, compared to the control rehabilitation group that did not use this application.

It is known that technology and the artificial intelligence applications it brings have the potential to increase the quality and safety of healthcare services. However, there are some concerns about the lack of reliability regarding this technology, its inadequate quality and its readability levels that the public can understand. (Grippaudo et al., 2024; Gül et al., 2024; Şahin et al., 2024; Ömür Arça et al., 2024) According to the standards determined by the United States Department of Health and Human Services, the American Medical Association and the National Institutes of Health, patient education materials should have a readability grade of six or below. (Gül et al., 2024; Ömür Arça et al., 2024; Erkin, Hanci & Ozduran, 2023; Özduran & Hanci, 2022; Ozduran & Büyükçoban, 2022).

There are increasing number of studies in the literature evaluating the reliability, readability and quality of AI chatbots on low back pathologies. Coraci et al, (2023) studied the development of medical questionnaires for low back pain in ChatGPT. As a result, although they found a significant correlation between other low back pain surveys and the ChatGPT survey, they stated that the power of this artificial intelligence chatbot was limited. Shrestha et al., (2024) studied the performance of ChatGPT in producing a clinical guideline in the diagnosis and treatment of LBP. They found that although ChatGPT provides an adequate clinical guideline

Commented [JA3]: Delete the full stop. There is another one after the reference

Commented [JA4]: Again the reference before the full stop

Commented [JA5]: Delete comma

recommendation, it tends to incorrectly recommend evidence. Yilmaz Muluk & Olcucu (2024) examined the effectiveness of ChatGPT-3.5 and GoogleBard in detecting Red Flags of LBP. They found that these AI chatbots showed strong performance but contained irrelevant content and showed low sensitivity. Gianola et al., (2024) evaluated the performance of ChatGPT in making informed decisions for lumbosacral radicular pain compared to clinical practice guidelines. They found that ChatGPT performed poorly in terms of internal consistency and accuracy of the generated indications compared to clinical practice guideline recommendations for lumbosacral radicular pain. Nian et al., (2024) searched patient education materials on Lumbar Spinal Fusion and Laminectomy on ChatGPT and Google. They found that ChatGPT responses were longer (340.0 vs. 159.3 words) and had lower readability (Flesch- Kincaid grade level: 11.6 vs. 8.8, Flesch Reading Ease score: 34.0 vs. 58.2) compared to Google. The authors noted that although ChatGPT was able to produce relatively accurate responses to certain questions, its role can be seen as a complement to consultation with a physician and should be used with caution until its functionality is validated (Nian et al., 2024).

Artificial intelligence chatbots have been studied not only on low back pain-related issues but also on different medical subjects, and impressive results have been obtained. Gül et al. (2024) found that the readability levels of Bard, ChatGPT and Perplexity responses to 100 questions related to subdural hematoma were higher than the recommended 6th grade level. They reported that although AI chatbots offer the opportunity to improve health outcomes and patient satisfaction, they are not sufficient in terms of readability. Şahin et al., (2024), evaluated the responses of 5 different artificial intelligence chatbots named Bard, ChatGPT, Ernie, Bing and Copilot to questions about erectile dysfunction in their study. They found that the AI chatbots that requires a high level of training to be understood is ChatGPT and the chatbot with the easiest readability is BARD. They reported that new AI chatbots to be developed in the future can provide more advanced counseling to patients if their understandability is easier (Şahin et al., 2024).

The increase in online sources of information raises concerns about which sources of information patients can trust and take into account. As mentioned above in the literature, many popular AI chatbots have been discussed in different studies and the information they contain has been analyzed in depth. However, there were no comparative studies of the 3 most popular AI chatbots on LBP in the literature. In line with this information, this study aimed to examine

Commented [JA6]: Delete comma

Commented [JA7]: Delete comma

Commented [JA8]: Delete comma

Commented [JA9]: Delete comma

the quality, reliability, and readability of the responses given by 3 different AI chatbots (ChatGPT, Gemini, and Perplexity) to frequently asked keywords about LBP.

MATERIAL & METHODS

Ethics Committee Permission

This cross-sectional study was prepared after receiving ethics committee approval (Cumhuriyet University Ethics Committee, Ethics Committee No: 2024/05-27, Date: 16.05.2024)

Research Procedure

The research was initiated by deleting all data sets belonging to personal internet browsers. After logging out of Google accounts, the research was continued by activating Google Incognito mode. The most frequently searched keywords related to low back pain were tried to be reached on May 29, 2024 in the Google Trends (https://trends.google.com/) search engine (Hershenhous et al, 2024). The search criteria were created by selecting health subheadings from all over the world from 2004 to the present. In the results section, the "most relevant" keywords were marked. As a result of the Google trend search, the 25 most frequently searched keywords with different categories were recorded. Geographical areas of interest were classified and recorded on the basis of subregions.

The keywords obtained were entered separately in English to ChatGPT, Gemini and Perplexity AI chatbots, which are freely accessible to everyone (Gül et al., 2024; Currie, Robbie & Tually, 2023). In order to prevent possible bias that could be created by the sequential processing of keywords in the answers given by the programs, the study was designed by providing input from different users (EO, VH) for each keyword. A different user was not assigned for each keyword and fake accounts were not used. The answers were recorded in the database so that they could be examined in terms of readability, reliability and quality. The keywords and responses from each AI chatbots are available from the web archive located at: https://archive.org/details/assessing-the-readability-quality-and-reliability-of-responses-produced-by-chat- Instead of ChatGPT Plus, the study was carried out using the GPT-4o version in the ChatGPT Free AI chatbot, which is free to everyone. In our study, AI chatbots that are freely accessible and accessible to people with low socioeconomic status were used (Gül et al., 2024; Ömür Arça et al., 2024).

Reliability Analysis

The reliability level of the answers was determined in the analysis based on "The Journal of the American Medical Association (JAMA) Benchmark". In order for a study to meet the JAMA criteria, it must meet four basic criteria such as authorship, currency, disclosure, and attribution. In the evaluation made according to the JAMA criteria, zero or one point is given for each criterion and these points are added up to form a general evaluation of the study between 0 and 4 points. Higher scores indicate that the study is more reliable, while lower scores indicate that it is less reliable (Kara et al., 2024; Ozduran & Hanci, 2023).

Another reliability scale used in our study is the Modified DISCERN scale. In this scale consisting of five criteria, if the required criterion is found, it is represented by 1 point, if not, it is represented by 0 points. Studies evaluated on a 5-point scale are considered more reliable as they receive higher scores (Erkin, Hanci & Ozduran, 2023a).

The questions in the scale can be listed as follows: "Is the literature review based on up-to-date and accurate sources?", "Are additional information sources listed for patient reference?", "Does the study address discussions in its field?" "Is the text clear and understandable?", "Is the information provided balanced and unbiased?". (Erkin, Hanci & Ozduran, 2023b) The validity and reliability of the JAMA and DISCERN scales have been evaluated (Silberg, Lundberg & Musacchio, 1997; Charnock et al., 1999). According to literature DISCERN instrument can be applied by experienced users and providers of health information to discriminate between publications of high and low quality. Chance corrected agreement (weighted kappa) for the overall rating was found kappa = 0.53 (95% CI kappa = 0.48 to kappa = 0.59) among the expert panel.The instrument will also be of benefit to patients, though its use will be improved by training. (Charnock et al., 1999).

Quality Analysis

Global Quality Score (GQS) is a system that evaluates the quality of online health information out of 5. 1 point indicates the lowest quality, 5 points the highest quality. According to this system, a source with a score of 1 is not of any quality for patients, while a source with a score of 5 is considered very high quality. In addition; 2 points: low quality, limited use. 3 points: medium quality, limited benefit. 4 points: good quality, useful (Gunduz et al., 2024). The Ensuring Quality Information for Patients (EQIP) is a tool that evaluates the quality and clarity of the relevant medical text. The 20 questions in this tool are answered as 'yes', 'partially' or 'no'. According to the answers given, the quality of the information is determined with a score

Commented [JA10]: Why the “” signs

between 0 and 100. When answering the 20 questions in the scale, 1 point is given to the "yes" answer, 0.5 to the partially answer and 0 to the no answer. The obtained scores are added and divided by 20, and then those that do not apply are removed and multiplied by 100 ((X of Yes*1) + (Y of Partly*0.5) + (Z of No*0))/20 - (Q of does not apply))] *100 = % score) (Ladhar et al., 2022). According to the EQIP tool results, those between “0%-25” are evaluated as “severe problems with quality”, those between “26%-50” are evaluated as “serious problems with quality”, “51%-75% are evaluated as “good quality with minor problems”, and “76% to 100% results are evaluated as “well written” (Moult, Franck & Brady, 2004). Reliability and validity assessments were made for the GQS and EQIP survey (Moult, Franck & Brady, 2004; Bernard et al., 2007). For example, The EQIP tool demonstrated strong validity, reliability, and utility in assessing the quality of a wide range of health information materials when employed by healthcare professionals and patient information managers. The internal consistency of the scale, as measured by Cronbach's alpha, was 0.80. Inter-rater reliability was also satisfactory, with a mean agreement of 0.60. (Moult, Franck & Brady, 2004).

Commented [JA11]: delete full stop

Readability Assessment

The responses given by AI chatbots to keywords were evaluated on two different websites that have the feature of calculating readability scores (http://readabilityformulas.com/ - (Calculator 1, https://www.online-utility.org/english/readability_test_and_improve.jsp - (Calculator 2)). The formulas used in text readability were Linsear Write (LW), Coleman-Liau Readability Index (CLI), Automated Readability Index (ARI), Simple Measure of Gobbledygook (SMOG), Gunning Fog Readability (GFOG), The Flesch Reading Ease Score (FRES) and Flesch Kincaid Grade Level (FKGL) (Gül et al., 2024; Özduran & Hanci, 2022; Hanc1 et al., 2024). Details on how readability was calculated with the formulas are given in Table 1. Final readability scores were recorded as median (minimum-maximum). The obtained responses were based on the sixth-grade It was analyzed with the readability level. Accordingly, the accepted average readability level is 80.0 for FRES and 6 for the other 6 formulas (Gül et al., 2024). The readability, quality and reliability level evaluation of the texts generated by artificial intelligence was carried out by two senior authors (EÖ and VH) with experience in the field of pain and the arithmetic average of the scores obtained in all 3 categories was taken.

Statistical Analysis

Data analysis was performed using SPSS Windows version 24.0 (SPSS Inc., USA). Frequency data are presented as numbers (n) and percentages (%), while continuous data are shown as

medians (minimum-maximum). Fisher's exact test and the Chi-square test were used to compare frequency variables, while the Mann-Whitney U and Wilcoxon tests were employed to compare continuous variables. To assess the consistency of the calculators, intraclass correlation coefficient (ICC) analysis was performed for each formula. Statistical significance was set at $p < 0.05$.

RESULTS

The first 3 keywords detected as a result of the Google Trend search were “Lower Back Pain”, “ICD 10 Low Back Pain”, and “Low Back Pain Symptoms”. The keyword “Pain in low back” was removed from the analysis because the keyword “Lower back pain” was present. The keywords “Low Back Pain ICD” and “Low Back Pain ICD 10 code” were removed from the analysis because the keyword “Low Back Pain ICD 10” was present. The keyword “Chronic Back Pain” was removed from the analysis because the keyword “Chronic Low Back Pain” was present. The keywords “Low back exercises” and “Back exercises” were removed from the analysis because the keyword “Low back pain exercises” was present. The keyword “Low Back Pain Kidney” was removed from the analysis because the keyword “Kidney Pain” was present. The keyword “Low Back Pain cause” was removed from the analysis because the keyword “Low Back Pain causes” was present. The keyword “what is Lower Back Pain” was removed from the analysis because the keyword “Lower back pain” was present. The keywords “Icd 10”, “lowback hip pain”, “hip pain” were removed because they were unrelated to the topic or showed different anatomical localizations. The study was organized on 13 keywords in the final. The full list of keywords is presented in Table 2. The United States, Puerto Rico, and Canada were determined to be the 3 countries with the highest searches for low back pain, respectively.

Keywords related to low back pain were entered into ChatGPT-4o, Perplexity and Google Gemini AI chatbots. The final readability scores of the responses given by these AIs were measured with two different programs, Calculator 1 and 2. The readability levels of the texts were evaluated by comparing them with the ability of a 6th grade reader to understand the text. The relevant results are given in Table 3, Table 4, Table 5 and Table 6.

Assessment of readability across the three groups, utilizing average scores from Calculators 1 and 2

In the analysis, the readability results of ChatGPT's answers were found to be more difficult in Calculator 1 for the GFOG, FKGL, CLI and ARI readability formulas, and more difficult in

Calculator 2 for the FRES, and SMOG readability formulas. Additionally, the readability results of Gemini's and Perplexity's answers were found to be more difficult in Calculator 1 for the GFOG and ARI readability formulas, and more difficult in Calculator 2 for the FRES, FKGL, CLI and SMOG readability formulas. When assessing the readability of responses among all three groups by averaging the outcomes from Calculator 1 and 2, significant differences emerged between specific groups. A significant difference ($p = 0.004$) was detected between ChatGPT-4o and Gemini in the CLI readability formula, but not in the other formulas. A significant difference was found between ChatGPT-4o and Perplexity in FOG, FKGL, SMOG and ARI readability formulas ($p < 0.001$). A significant difference was found between Gemini and Perplextiy in all readability formulas ($p < 0.05$) (Table 5). Based on the readability assessments, all readability metrics, excluding GFOG and ARI, are arranged in a hierarchy of readability from easiest to most difficult: Google Gemini, ChatGPT-4o, and Perplexity. Nonetheless, as per the GFOG and ARI readability metric, the order varies slightly: ChatGPT-4o, Google Gemini, Perplexity (Table 5).

Assessing ChatGPT, Gemini, and Perplexity responses based on the suggested reading level for sixth graders

When the median readability scores of all responses were compared to the sixth-grade reading level, a statistically significant difference was observed for all metrics ($p < 0.001$). Importantly, the readability of the responses exceeded the sixth-grade standard across all metrics. Similarly, statistically significant results were found when comparing the outcomes from Calculator 1 and Calculator 2, as well as the combined average of both calculators ($p < 0.001$) (Tables 3, 4, and 5).

Reliability and Quality Assessment

Perplexity's answers achieved the top EQIP, JAMA, modified DISCERN and GQS scores ($P < 0.001$) (Table 6). According to these results, it can be said that Perplexity offers more reliable and quality data to its users.

Intraclass correlation coefficients (ICC)

GFOG,FRES,CL, FKGL, ARI and SMOG scores were computed using two different calculators (https://www.online-utility.org/english/readability_test_and_improve.jsp, https://readabilityformulas.com/free-readability-formula-tests.php).

ICC for ChatGPT

The intraclass correlation coefficient for FRES was 0.942, for FKGL was 0.876, for GFOG was 0.827, for CL was 0.951, for ARI was 0.827 and for SMOG was 0.852.

ICC for Gemini

The intraclass correlation coefficient for FRES was 0.973, for KFGL was 0.985, for GFOG was 0.984, for CL was 0.972, for ARI was 0.978 and for SMOG was 0.976.

ICC for Perplexity

The intraclass correlation coefficient for FRES was 0.930, for FKGL was 0.961, for GFOG was 0.955, for CL was 0.966, for ARI was 0.943 and for SMOG was 0.939.

ICC for GQS, JAMA, mDISCERN and EQIP

The intraclass correlation coefficients were 0.915 for GQS, 0.981 for JAMA, 0.898 for mDISCERN 0.984 for EQIP.

According to these results, it can be said that there is a very strong correlation between the readability scores given by both calculators in our study and the quality and reliability survey answers given by both authors.

DISCUSSION

This study evaluated the quality, readability, and reliability of responses to frequently asked keywords about LBP provided by Perplexity, Gemini, and ChatGPT AI chatbots. Responses with reading levels higher than the 6th-grade reading level recommended by the U.S. Department of Health and Human Services and the National Institutes of Health were detected in all 3 AI chatbots. In addition, it was determined that Perplexity's responses received higher results in reliability and quality analysis than other chatbots. To our knowledge, this study evaluated the information quality, reliability, and readability levels of responses to frequently asked keywords about LBP generated by 3 different popular AI chatbots and represents the first and pioneering research effort on this topic.

A key factor in understanding patient education materials is their readability. Complex and long sentences are known to undermine the reader's confidence, making it difficult to learn health-related written texts. Furthermore, it has been shown that having 8 to 10 words in a sentence facilitates the readability of health-related information (Gül et al., 2024). More readable texts

will help create better health literacy, increasing patient compliance, reducing emergency care visits, and shortening hospital stays (Hancı et al., 2024).

In the literature, online information on LBP has been studied and shown to have readability levels that are more difficult than recommended. In a study on acute LBP, only 3 of 22 websites providing online information provided an acceptable readability score, while the readability levels of the other websites were reported to exceed the recommended level for the average person to understand (Hendrick et al, 2012). Another study investigated 72 websites on failed back spinal surgery and found that they were of low quality and content due to low JAMA and DISCERN scores (Guo et al, 2019). Studies frequently mentioned the detrimental effects of difficult readability and low quality and low reliability information on public health (Hendrick et al, 2012; Guo et al, 2019). AI chatbots have become a platform where a significant portion of patients seek answers to their medical questions due to their accessibility and ability to provide personalized answers. Some clinicians also consider these chatbots as a tool with the potential to improve patient education due to their broad knowledge base and ability to produce consistent and original answers (Ömür Arça et al, 2024). Therefore, our study aimed to evaluate the readability, quality and reliability parameters of the answers given by popular AI chatbots to keywords about LBP, not online websites.

There are also studies in the literature on the readability of the information provided by AI chatbots on LBP-related issues. Scaff et al. (2024), examined the answers to 30 questions about LBP to 4 different AI chatbots, namely ChatGPT 3.5, Bing, Bard and ChatGPT 4.0, and found that the answers were poor and could negatively impact patient understanding and behavior. The authors emphasized that poorly readable texts can challenge patients, potentially leading to misinformation, inappropriate care, and worsened health outcomes. Nian et al. (2024), found that ChatGPT's answers to questions about lumbar spinal fusion and laminectomy were accurate but not specific enough and had readability appropriate for a slightly above average health literacy level. Each response resulted in a recommendation for further consultation with a healthcare provider, emphasizing to patients the value of physician consultation. There are studies in the literature showing that AI chatbots provide answers with high readability scores on different topics (Gül et al., 2024; Şahin et al., 2024).

In our study, it was determined similar to literature that the responses given by the AI chatbots to the most frequent keywords about LBP had readability levels higher than the 6th grade level

recommended by the United States Department of Health and Human Services, the American Medical Association, and the National Institutes of Health. The easiest readability was determined in Google Gemini, while the most difficult readability was determined in Perplexity. Developing AI chatbots with appropriate readability will help patients trying to access health-related online information to reach more understandable information.

In our study, not only the readability of AI chatbots but also their quality and reliability were tested. In the study examining the responses given by 4 different chatbots, ChatGPT, BARD, Gemini and Copilot, to questions about palliative care in the literature, it was reported that none of the ChatGPT responses met the JAMA criteria. In addition, it was determined that mDISCERN and JAMA scores were the highest in Perplexity, and GQS scores were the highest in Gemini (Hancı et al., 2024). In the study conducted by Casciato et al. (2024) on the information given by AI chatbots on foot and ankle surgery, they determined that they had low reliability and accuracy due to low JAMA and DISCERN scores. There are some studies in the literature showing that Perplexity produces answers with high DISCERN and JAMA scores (Gül et al., 2024, Hancı et al., 2024). In our study, high mDISCERN, JAMA, GQS and EQIP scores were detected in Perplexity. These scores were significantly lower in other AI chatbots. In the future, new chatbots or new versions of existing chatbots can be produced, and these adjustments can be made to provide higher quality and more reliable information to those requesting online health-related information. Additionaly, it is an undeniable fact that none of the information provided by these chatbots can replace a face-to-face medical consultation.

Potential Implications for Clinical Practice and Future Studies

This study highlighted the potential and limitations of AI chatbots' responses regarding LBP. Although it is emphasized that the answers obtained help to provide significant gains to public health by providing reliable and quality content, the fact that answers with high readability scores for patients with average health literacy may cause patients to make incorrect or incompletely informed health decisions and receive inappropriate or delayed health care. Additionally, this study can provide inspiration for future studies on improving the algorithms and responses of AI chatbots and provide guidance for possible policies regarding AI's appropriate information delivery on patient education and health literacy.

Limitations of the study

We can list the limitations in our study as follows. The study we planned using the 25 most popular keywords offered by Google regarding low back pain can be made more comprehensive with studies to be produced using more keywords in the future. Another limitation is the presence of only English-language keywords in our study. In addition, studies to be conducted with other chatbots other than Gemini, ChatGPT and Perplexity will reveal the functioning of different artificial intelligence models. In our study, we evaluated the responses given by chatbots to the keywords detected on a day in May 2024. This situation shows that different keywords that can be obtained on another date may yield different study results.

Strength of the study

Our study is the first to demonstrate not only the readability but also the quality and reliability of AI chatbots on LBP. Unlike many other study methodologies, the fact that we evaluated the responses of multiple popular AI chatbots, not just a single AI chatbot, can be considered another strength of the study.

CONCLUSION

AI chatbots such as ChatGPT, Perplexity and Gemini are increasingly performing well in providing medical information. This may provide an opportunity to raise awareness and improve patient satisfaction on issues related to LBP. Despite this, there are still some concerns about the readability, quality and reliability assessment results of AI chatbots. In our study, it has been determined that the answers given by AI chatbots to keywords about LBP are difficult to read and have low reliability and quality assessment. In the future, the information provided in AI chatbots will be presented through an expert team review and texts with appropriate and understandable readability will positively affect public health.

REFERENCES

Ács, P., Betlehem, J., Oláh, A., Bergier, B., Morvay-Sey, K., Makai, A., Prémusz, V. Cross-cultural adaptation and validation of the Global Physical Activity Questionnaire among healthy Hungarian adults. BMC public health, 2020;20(Suppl 1), 1056. DOI 10.1186/s12889-020-08477-z

Bagcier, F., Yurdakul, O. V., Ozduran, E. Top 100 cited articles on ankylosing spondylitis. Reumatismo, 2021;72(4), 218–227. DOI 10.4081/reumatismo.2020.1325

Bernard, A., Langille, M., Hughes, S., Rose, C., Leddin, D., Veldhuyzen van Zanten, S. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. The American journal of gastroenterology, 2007;102(9), 2070–2077. DOI 10.1111/j.1572-0241.2007.01325.x

Casciato, D., Mateen, S., Cooperman, S., Pesavento, D., Brandao, R. A. Evaluation of Online AI-Generated Foot and Ankle Surgery Information. The Journal of foot and ankle surgery : official publication of the American College of Foot and Ankle Surgeons, 2024;S10672516(24)00143-1. Advance online publication. DOI 10.1053/j.jfas.2024.06.009

Charnock, D., Shepperd, S., Needham, G., Gann, R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. Journal of epidemiology and community health, 1999;53(2), 105–111. DOI 10.1136/jech.53.2.105

Coraci, D., Maccarone, M. C., Regazzo, G., Accordi, G., Papathanasiou, J. V., & Masiero, S. ChatGPT in the development of medical questionnaires. The example of the low back pain. European journal of translational myology, 2023;33(4), 12114. DOI 10.4081/ejtm.2023.12114

Currie G, Robbie S, Tually P. ChatGPT and Patient Information in Nuclear Medicine: GPT- 3.5 Versus GPT-4.J Nucl Med Technol. 2023;5;51(4):307-313. DOI 10.2967/jnmt.123.266151

DePalma M. G. Red flags of low back pain. JAAPA : official journal of the American Academy of Physician Assistants, 2020;33(8), 8–11. DOI 10.1097/01.JAA.0000684112.91641.4c

Do, K., Kawana, E., Vachirakorntong, B., Do, J., Seibel, R. The use of artificial intelligence in treating chronic back pain. The Korean journal of pain, 2023;36(4), 478–480. DOI 10.3344/kjp.23239

Erkin Y, Hanci V, Ozduran E. Evaluating the readability, quality and reliability of online patient education materials on transcutaneuous electrical nerve stimulation (TENS). Medicine (Baltimore). 2023;102(16):e33529. DOI 10.1097/MD.0000000000033529

Erkin Y, Hanci V, Ozduran E. Evaluation of the reliability and quality of YouTube videos as a source of information for transcutaneous electrical nerve stimulation. PeerJ. 2023a;21;11:e15412. DOI 10.7717/peerj.15412.

Erkin, Y., Hanci, V., Ozduran, E. Evaluating the readability, quality and reliability of online patient education materials on transcutaneuous electrical nerve stimulation (TENS). Medicine, 2023b;102(16), e33529. DOI 10.1097/MD.0000000000033529

Gianola, S., Bargeri, S., Castellini, G., Cook, C., Palese, A., Pillastrini, P., Salvalaggio, S., Turolla, A., Rossettini, G. (Performance of ChatGPT Compared to Clinical Practice Guidelines in Making Informed Decisions for Lumbosacral Radicular Pain: A Cross- sectional Study. The Journal of orthopaedic and sports physical therapy, 2024;54(3), 222– 228. DOI 10.2519/jospt.2024.12151

Grippaudo, F. R., Nigrelli, S., Patrignani, A., Ribuffo, D. Quality of the Information provided by ChatGPT for Patients in Breast Plastic Surgery: Are we already in the future?. JPRAS open, 2024;40, 99–105. DOI 10.1016/j.jpra.2024.02.001

Gunduz, M. E., Matis, G. K., Ozduran, E., Hanci, V. Evaluating the Readability, Quality, and Reliability of Online Patient Education Materials on Spinal Cord Stimulation. Turkish neurosurgery, 2024;34(4), 588–599. DOI 10.5137/1019-5149.JTN.42973-22.3

Guo, W. J., Wang, W. K., Xu, D., Qiao, Z., Shi, Y. L., Luo, P. Evaluating the Quality, Content, and Readability of Online Resources for Failed Back Spinal Surgery. Spine, 2019;44(7), 494–502. DOI 10.1097/BRS.0000000000002870

Gül, Ş., Erdemir, İ., Hanci, V., Aydoğmuş, E., Erkoç, Y. S. How artificial intelligence can provide information about subdural hematoma: Assessment of readability, reliability, and quality of ChatGPT, BARD, and perplexity responses. Medicine, 2024;103(18), e38009. DOI 10.1097/MD.0000000000038009

Hancı, V., Ergün, B., Gül, Ş., Uzun, Ö., Erdemir, İ., Hancı, F. B Assessment of readability, reliability, and quality of ChatGPT®, BARD®, Gemini®, Copilot®, Perplexity® responses on palliative care. Medicine, 2024;103(33), e39305. DOI 10.1097/MD.0000000000039305

Hartmann, R., Avermann, F., Zalpour, C., Griefahn, A. Impact of an AI app-based exercise program for people with low back pain compared to standard care: A longitudinal cohort-study. Health science reports, 2023;6(1), e1060. DOI 10.1002/hsr2.1060

Hartvigsen, J., Hancock, M. J., Kongsted, A., Louw, Q., Ferreira, M. L., Genevay, S., Hoy, D., Karppinen, J., Pransky, G., Sieper, J., Smeets, R. J., Underwood, M., Lancet Low Back Pain Series Working Group What low back pain is and why we need to pay attention. Lancet (London, England), 2018;391(10137), 2356–2367. DOI 10.1016/S0140-6736(18)30480-X

Hemmer C. R. Evaluation and Treatment of Low Back Pain in Adult Patients. Orthopedic nursing, 2021;40(6), 336–342. DOI 10.1097/NOR.0000000000000804

Hendrick, P. A., Ahmed, O. H., Bankier, S. S., Chan, T. J., Crawford, S. A., Ryder, C. R., Welsh, L. J., Schneiders, A. G. Acute low back pain information online: an evaluation of quality, content accuracy and readability of related websites. Manual therapy, 2012;17(4),318–324. DOI 10.1016/j.math.2012.02.019

Hershenhouse, J. S., Mokhtar, D., Eppler, M. B., Rodler, S., Storino Ramacciotti, L., Ganjavi, C., Hom, B., Davis, R. J., Tran, J., Russo, G. I., Cocci, A., Abreu, A., Gill, I., Desai, M., Cacciamani, G. E. Accuracy, readability, and understandability of large language models for prostate cancer information to the public. Prostate cancer and prostatic diseases, 2024;10.1038/s41391-024-00826-y. DOI 10.1038/s41391-024-00826-y Hodges, P. W., Setchell, J., Nielsen, M. An Internet-Based Consumer Resource for People with Low Back Pain (MyBackPain): Development and Evaluation. JMIR rehabilitation and assistive technologies, 2020;7(1), e16101. DOI 10.2196/16101

Járomi, M., Szilágyi, B., Velényi, A., Leidecker, E., Raposa, B. L., Hock, M., Baumann, P., Ács, P., Makai, A. Assessment of health-related quality of life and patient's knowledge in chronic non-specific low back pain. BMC public health, 2021;21(Suppl 1), 1479. DOI 10.1186/s12889-020-09506-7

Kara, M., Ozduran, E., Mercan Kara, M., Hanci, V., Erkin, Y. Assessing the quality and reliability of YouTube videos as a source of information on inflammatory back pain. PeerJ, 2024;12, e17215. DOI 10.7717/peerj.17215

Ladhar, S., Koshman, S. L., Yang, F., Turgeon, R. Evaluation of Online Written Medication Educational Resources for People Living With Heart Failure. CJC open, 2022;4(10), 858-865. DOI 10.1016/j.cjco.2022.07.004

Moult, B., Franck, L. S., Brady, H. Ensuring quality information for patients: development and preliminary validation of a new instrument to improve the quality of written health care information. Health expectations : an international journal of public participation in health care and health policy, 2004;7(2), 165–175. DOI 10.1111/j.1369-7625.2004.00273.x

Nian, P. P., Saleet, J., Magruder, M., Wellington, I. J., Choueka, J., Houten, J. K., Saleh, A., Razi, A. E., Ng, M. K. ChatGPT as a Source of Patient Information for Lumbar Spinal Fusion and Laminectomy: A Comparative Analysis Against Google Web Search. Clinical spine surgery, 2024;10.1097/BSD.0000000000001582. Advance online publication. DOI 10.1097/BSD.0000000000001582

Nolet, P. S., Kristman, V. L., Côté, P., Carroll, L. J., Cassidy, J. D. The association between a lifetime history of low back injury in a motor vehicle collision and future low back pain: a population-based cohort study. European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society, 2018;27(1), 136–144. DOI 10.1007/s00586-017-5090-y

Ozduran, E., Büyükçoban, S. Evaluating the readability, quality and reliability of online patient education materials on post-covid pain. PeerJ, 2022;10, e13686. DOI 10.7717/peerj.13686

Ozduran, E., Hanci, V. Youtube as a source of information about stroke rehabilitation during the COVID-19 pandemic. Neurology Asia, 2023;28(4). DOI 10.54029/2023kif

Ömür Arça, D., Erdemir, İ., Kara, F, Shermatov, N., Odacioğlu, M., İbişoğlu, E, Hanci, FB, Sağiroğlu, G., Hanci, V. Assessing the readability, reliability, and quality of artificial intelligence chatbot responses to the 100 most searched queries about cardiopulmonary

resuscitation: An observational study. Medicine 2024;103(22):p e38352. DOI 540 10.1097/MD.0000000000038352

Özduran, E., Hanci, V. Evaluating the readability, quality and reliability of online information on Behçet's disease. Reumatismo, 2022;74(2),10.4081/reumatismo.2022.1495. 543 DOI 10.4081/reumatismo.2022.1495

Scaff, S. P. S., Reis, F. J. J., Ferreira, G. E., Jacob, M. F., Saragiotto, B. T. Assessing the performance of AI chatbots in answering patients' common questions about low back pain. Annals of the rheumatic diseases, 2024;ard-2024-226202.. DOI 10.1136/ard-2024-226202

Shrestha, N., Shen, Z., Zaidat, B., Duey, A. H., Tang, J. E., Ahmed, W., Hoang, T., Restrepo Mejia, M., Rajjoub, R., Markowitz, J. S., Kim, J. S., Cho, S. K. Performance of ChatGPT on NASS Clinical Guidelines for the Diagnosis and Treatment of Low Back Pain: A Comparison Study. Spine, 2024;49(9), 640–651. DOI 10.1097/BRS.0000000000004915

Silberg, W. M., Lundberg, G. D., Musacchio, R. A. Assessing, controlling, and assuring the quality of medical information on the Internet: Caveant lector et viewor--Let the reader and viewer beware. JAMA, 1997;277(15), 1244–1245.

Şahin, M. F., Ateş, H., Keleş, A., Özcan, R., Doğan, Ç., Akgül, M., Yazıcı, C. M. Responses of Five Different Artificial Intelligence Chatbots to the Top Searched Queries About Erectile Dysfunction: A Comparative Analysis. Journal of medical systems, 2024;48(1), 38. DOI 10.1007/s10916-024-02056-0

Verhagen, A. P., Downie, A., Popal, N., Maher, C., Koes, B. W. Red flags presented in current low back pain guidelines: a review. European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society, 2016; 25(9), 2788–2802. DOI 10.1007/s00586-016-4684-0

Yilmaz Muluk, S., Olcucu, N. Comparative Analysis of Artificial Intelligence Platforms: ChatGPT-3.5 and GoogleBard in Identifying Red Flags of Low Back Pain. Cureus, 2024;16(7), e63580. DOI 10.7759/cureus.63580