Original paper

Reumatologia 2025; 63, 5: 313–320
DOI: https://doi.org/10.5114/reum/207526

Evaluating ChatGPT-4 for rheumatology patient education: a comparative analysis of readability, reliability, and similarity to the American College of Rheumatology's fact sheets

Yakup Erden[1], Mustafa Hüseyin Temel[2], Fatih Bağcıer[3]

[1]Physical Medicine and Rehabilitation Clinic, İzzet Baysal Physical Medicine and Rehabilitation Training and Research Hospital, Bolu, Turkey
[2]Physical Medicine and Rehabilitation Clinic, University of Health Sciences, Sultan 2. Abdulhamid Han Training and Research Hospital, Istanbul, Turkey
[3]Physical Medicine and Rehabilitation Clinic, Başakşehir Çam and Sakura City Hospital, Istanbul, Turkey

Abstract

Introduction: This study aimed to evaluate the readability, quality, reliability, similarity, and length of texts generated by ChatGPT on common rheumatic diseases and compare their content with American College of Rheumatology (ACR) patient education fact sheets.

Material and methods: Fifteen common rheumatic diseases were included based on the ACR fact sheets. Questions about disease characteristics, symptoms, treatments, and lifestyle recommendations were generated based on ACR content and input into ChatGPT-4 for comparison. Readability was assessed using the Flesch-Kincaid Grade Level (FKGL), Flesch Reading Ease (FRE), and the Simple Measure of Gobbledygook (SMOG) index. Quality and reliability were evaluated using the DISCERN questionnaire and the Ensuring Quality Information for Patients (EQIP) tool. Text similarity was measured using cosine similarity, and word count was obtained using Microsoft Word.

Results: ChatGPT-generated texts had significantly higher FKGL scores (14.3 vs. 12.7; $p = 0.007$) and SMOG scores ($p < 0.001$), indicating greater linguistic complexity. They also had lower FRE scores (35.8 vs. 43.7; $p < 0.001$). The mean DISCERN score for ChatGPT was significantly lower than for ACR fact sheets (46 vs. 52; $p < 0.001$), suggesting reduced reliability. However, no significant difference was found in EQIP quality scores ($p = 0.744$). Cosine similarity between ChatGPT and ACR texts averaged 0.69 (range: 0.57–0.76), indicating moderate content overlap. ChatGPT texts were more than twice as long, with a median word count of 1,109 compared to 450 for ACR materials ($p < 0.001$).

Conclusions: Despite the moderate similarity, ChatGPT-generated texts on rheumatic diseases were more complex, less reliable, and longer than ACR fact sheets. These findings highlight the need for improvements in artificial intelligence-driven healthcare tools to ensure readability, accuracy, and reliability, making them more aligned with expert-reviewed resources.

Key words: ChatGPT, rheumatic diseases, American College of Rheumatology, health information.

Introduction

Rheumatic diseases (RDs) encompass a range of chronic, inflammatory, and progressive disorders, including rheumatoid arthritis, spondyloarthropathies, Sjögren's disease, systemic lupus erythematosus, scleroderma, and dermatomyositis. They can significantly impact patient health, frequently resulting in disability and a notable decline in quality of life [1]. In addition to their effects on individuals, these diseases impose a significant burden on healthcare systems and society in general [2, 3].

The application of artificial intelligence (AI) in healthcare, particularly for managing chronic medical conditions such as RDs, is an expanding area of research and development. Individuals with chronic illnesses frequently seek information and assistance to manage

Address for correspondence
Yakup Erden, Physical Medicine and Rehabilitation Clinic, İzzet Baysal Physical Medicine and Rehabilitation Training and Research Hospital, Orüs Street No. 59, 14020 Bolu/Turkey, e-mail: yakuperden@hotmail.com
Submitted: 31.01.2025; **Accepted:** 24.06.2025

their afflictions, and AI-based tools, such as conversational agents and health coaching systems, are increasingly being created to address these needs [4, 5]. The integration of AI into the healthcare domain is progressing rapidly, with ChatGPT emerging as a prominent exemplar of this technological advancement. ChatGPT has gained widespread popularity among numerous users due to its capacity to generate detailed and rapid responses and its accessibility. As a large-scale language model, it employs deep learning techniques based on a variant of the transformer architecture to produce human-like responses to text-based input [6].

There is an increasing acknowledgment of the capacity of AI chatbots to provide prompt, precise, and empathetic information to individuals managing chronic health issues [7]. Nevertheless, the use of these tools in medical domains, particularly rheumatology, remains a largely unexplored area. Research has highlighted the strengths and weaknesses of these technologies, indicating that they can produce reasonably accurate and empathetic responses akin to those provided by medical professionals. The importance of ongoing management and patient education in rheumatology cannot be overstated. It is crucial to deliver information that is clear and easy to comprehend. Nonetheless, there is a valid concern that these chatbots might disseminate outdated or inaccurate information, underscoring the importance of a comprehensive evaluation of these emerging technologies [8, 9]. Indeed, the European Alliance of Associations for Rheumatology has acknowledged the potential of leveraging big data to tackle rheumatic and musculoskeletal diseases. There is a strong emphasis on the necessity for further benchmarking studies to assess the effectiveness and reliability of AI-driven healthcare tools [10].

Considering all the background information provided, the objective of this study was to evaluate the quality, reliability, and readability of texts generated by ChatGPT on common RDs and compare the similarity and word count of these texts to the fact sheets created by the American College of Rheumatology (ACR) to bridge the gap in the literature and provide insights for future studies. The ACR was selected due to its prominent international presence and active leadership in global rheumatology collaboration, which support its visibility and influence in standard-setting efforts [11].

Material and methods

This study was conducted at the İzzet Baysal Physical Treatment and Rehabilitation Training and Research Hospital from August 20, 2024 to September 15, 2024. The study adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for cross-sectional studies to ensure transparent and standardized reporting [12].

Selection of diseases

Fifteen common RDs were identified from the fact sheets available on the official website of the ACR [13]:

- rheumatoid arthritis,
- systemic lupus erythematosus,
- spondyloarthritis,
- psoriatic arthritis,
- fibromyalgia,
- gout,
- Sjögren's disease,
- osteoarthritis,
- scleroderma,
- polymyalgia rheumatica,
- vasculitis,
- inflammatory myopathies,
- reactive arthritis,
- familial Mediterranean fever,
- Behçet's disease.

Fact sheets that included comprehensive information on etiology, common signs and symptoms, treatment options, and lifestyle recommendations were retained for analysis. Those missing any of these key domains were excluded to maintain a standardized framework for comparison with the texts produced by ChatGPT.

Data collection

All browsing data were cleared entirely before initiating the searches, and a new account was created to engage with ChatGPT as a precautionary step to avoid any browsing history bias. Each RD query was handled on distinct chat pages to maintain clarity and enhance the efficiency of the analytical process.

Questions about common RDs were developed explicitly for this study based on the health information in the ACR fact sheets [13] to compare the readability, quality, and similarity of the ACR fact sheets and the texts generated by ChatGPT-4. The prompts submitted to ChatGPT were systematically developed based on the structural and thematic organization of the ACR fact sheets, incorporating key domains such as disease etiology, clinical features, treatment modalities, and lifestyle recommendations. A single, standardized prompt was used in English for each condition, and no iterative regeneration or manual optimization was performed. This methodological approach was intended to simulate typical, real-world patient interactions with AI-driven tools, in which users generally input straightforward, natural-

language queries without applying advanced formatting or refinement techniques [14]. The study aimed to assess the model's baseline performance in generating patient-facing health information under ecologically valid conditions by avoiding prompt manipulation. Prompt engineering strategies were deliberately excluded, as they represent a specialized skill set not widely adopted by the general public or typical end-users seeking medical information [15].

The following standardized questions were used in the study to evaluate the content related to common RDs:

- "What is [disease]?"
- "What are the signs/symptoms of [disease]?"
- "What are common treatments for [disease]?"
- "What are tips for living with [disease]?"

The obtained texts were then evaluated using established metrics. The meanings and interpretations of the text evaluation metrics, along with their formulas, are provided in Table I.

Readability assessment

The Flesch-Kincaid Grade Level (FKGL), Flesch Reading Ease (FRE), and Simple Measure of Gobbledygook (SMOG) Index metrics were used to assess the readability of the content generated by AI chatbots. The FKGL is calculated by is calculated by multiplying the average sentence length (words per sentence) by 0.39, adding the average syllables per word multiplied by 11.8, and subtracting 15.59. A lower score signifies easier comprehension, while a higher one suggests greater linguistic complexity. The FRE measures document readability through calculations involving average sentence length multiplied by 1.015 and average number of syllables per word multiplied by 84.6; their difference is then subtracted from 206 [16]. The SMOG grade is calculated by multiplying 1.0430 by the square root of the total number of polysyllabic words multiplied by 30 divided by the total number of sentences, and then adding 3.1291 to the result. Finally, the result is rounded to the nearest whole number to determine the reading grade level. The score corresponds to a U.S. school grade level; a higher SMOG score indicates a more complex text [17].

Reliability and quality assessment

The DISCERN questionnaire, a validated tool developed to assist patients and information providers in evaluating the quality and reliability of the written content on treatment options, was used. The DISCERN tool is often used to assess the quality of health information based on criteria such as reliability, accuracy, and clarity. The minimum DISCERN score is 15; the maximum score is 75. DISCERN scores are categorized as follows: a score of 63 to 75 indicates excellent, a score of 51 to 62 is good, a score of 39 to 50 is fair, a score of 27 to 38 is poor, and a score of 16 to 26 is very poor [18].

The Ensuring Quality Information for Patients (EQIP) tool was used to analyze the quality of the gathered texts. This tool assesses different aspects of the material, including coherence and writing quality. The questionnaire consists of 20 inquiries, with response options including "yes", "partly", "no", or "does not apply". The scoring approach entails the multiplication of the quantity of

Table I. Formulas and interpretations of text evaluation metrics

Metric	Formula	Interpretation
Flesch-Kincaid Grade Level	(total words/total sentences × 0.39) + (total syllables/total words × 11.8) – 15.59	A lower score indicates easier comprehension; a higher score indicates more complex text
Flesch Reading Ease	206 – (1.015 × average sentence length) – (84.6 × average syllables per word)	A higher score indicates easier readability; a lower score indicates harder readability
Simple Measure of Gobbledygook Index	1.0430 × square root (total polysyllabic words × [30 ÷ total sentences]) + 3.1291	A higher score means more complex text, equivalent to a higher U.S. school grade level
DISCERN Score	Rating reliability, accuracy, and clarity of health information (15–75)	63–75 = excellent, 51–62 = good, 39–50 = fair, 27–38 = poor, 16–26 = very poor
Ensuring Quality Information for Patients Score	Sum of responses (Yes = 1, Partly = 0.5, No = 0) divided by total applicable items, multiplied by 100	Higher scores (76–100%) indicate well-written, high-quality information. Lower scores indicate quality issues
Cosine similarity	Cosine of the angle between the 2 vectors representing the text using TF-IDF	A value closer to 1 indicates high similarity, while a value closer to 0 indicates low similarity
Word count	Total number of words in the text	Provides an idea of text length. More words may indicate more detailed content, but can also suggest verbosity

TF-IDF – term frequency – inverse document frequency.

"yes" responses by 1, "partly" responses by 0.5, and "no" responses by 0. The resultant values are aggregated, divided by the total quantity of items, and adjusted by removing the count of responses labeled as "does not apply". Resources with scores ranging from 0 to 25% were categorized as "severe quality issues", 26–50% as "serious quality issues", 51–75% as "good quality with minor issues", and 76–100% as "well written", indicating exceptional quality [19].

Two independent raters (Y.E. and M.H.T.) evaluated all texts using the DISCERN and EQIP tools. Both raters jointly reviewed the assessment criteria before scoring to ensure consistency. All evaluations were performed independently and blinded to the text source (ChatGPT or ACR). In scoring discrepancies, a third reviewer (F.B.) acted as an arbitrator to resolve disagreements and determine the final rating.

Similarity and text length assessment

Cosine similarity, a well-established and widely used metric in text analysis, was employed to quantify the textual similarity between the materials. This metric measures the cosine of the angle between 2 numerical vectors, thereby providing a means to assess the similarity between textual elements [20]. Specifically, the scikit-learn library for Python was used, wherein the text was first transformed into a numerical representation using the Term Frequency-Inverse Document Frequency technique. Subsequently, the "cosine_similarity" function was applied to compute the cosine similarity between the transformed textual elements. Cosine similarity ranges from 0 to 1, with 1 indicating identical vectors and 0 indicating no similarity [21].

The text's word count was determined using Microsoft Word (Microsoft Corporation, Redmond, WA, USA). The built-in word count tool, accessible via the "Review" tab, calculated the total number of words [22, 23].

All statistical analysis was conducted using SPSS version 27 (IBM, New York, USA). The data normality was assessed with the Shapiro-Wilk test. Given the relatively small sample sizes ($n < 30$ per group) and the limitations of normality testing under these conditions, non-parametric methods were selected to reduce the risk of assumption violations [24]. Continuous data are represented as mean ±standard deviation, median (min.–max.) for non-normally distributed data, and categorical data as frequency. Between-group differences were computed with the Kruskal-Wallis test. Any possible correlation was investigated using the Spearman correlation coefficient. Post-hoc analysis was performed using the Bonferroni test. The significance level was 0.05.

Bioethical standards

Ethical committee approval for this study was not sought as it did not include any procedures on human or animal data, and it was conducted using publicly available data.

Results

The Shapiro-Wilk test showed that FRE, FKGL, and SMOG scores were normally distributed ($p > 0.05$), while EQIP scores, DISCERN scores, and word count values were not ($p < 0.05$). Due to small sample sizes and non-normal distributions in several variables, non-parametric tests were applied [24].

The total word count, readability, reliability, and quality scores of the texts are summarized in Table II.

Readability

The mean SMOG score for all ACR fact sheets was 12.72 ±1.15, compared with 14.30 ±0.80 for ChatGPT-generated texts ($p < 0.001$). The mean FKGL score for ACR fact sheets was 11.4 ±1.42, compared with 12.57 ±1 for ChatGPT-generated texts. The mean FRE score for ACR fact sheets was 43.75 ±9.40, compared with 35.83 ±5.5 for ChatGPT-generated texts ($p < 0.001$).

Reliability and quality

The median DISCERN score for ACR fact sheets was 52 (min.–max.: 48–55), while the median score for ChatGPT-generated texts was 46 (min.–max.: 44–49; $p < 0.001$). The EQIP scores showed no significant difference between ACR fact sheets and ChatGPT-generated texts ($p = 0.744$).

Similarity and text length

The cosine similarity index values between ACR fact sheets and ChatGPT-generated texts ranged from 0.57 to 0.74, with an average of 0.69 ±0.05 (Table III). The median word count for ACR information pages was 450 (min.–max.: 361–553), compared to 1109 (min.–max.: 929–1,274) for ChatGPT-generated texts ($p < 0.001$).

Discussion

This study provides early empirical evidence that ChatGPT-generated texts on RDs are significantly less readable, longer, and less reliable than expert-authored materials such as ACR fact sheets. To our knowledge, this is the first study to benchmark ChatGPT's educational content in rheumatology against validated pa-

Table II. Comparative analysis of word count readability and quality of texts produced by ChatGPT vs. ACR

	ChatGPT			ACR			*p*
	Min.	Max.	Mean	Min.	Max.	Mean	
Flesch Kincaid Grade Level	10.58	13.92	12.57 ±1.0	8.56	14.22	11.14 ±1.42	0.007
Flesch Reading Ease Score	27.49	46.43	35.83 ±5.5	24.53	61.17	43.75 ±9.40	0.005
Simplified Measure of Gobbledygook Index	12.61	15.20	14.30 ±0.8	10.83	14.75	12.72 ±1.15	< 0.001
			Median			Median	
DISCERN Score	44	49	46	48	55	52	< 0.001
Ensuring Quality Information for Patients Tool	0.65	0.73	0.69	0.75	0.95	0.84	0.744
Total word count	929	1274	1,109	361	553	450	< 0.001

ACR – American College of Rheumatology.

tient resources, revealing both the potential and limitations of AI-generated health information.

A major issue identified was the readability gap. While ACR materials adhered to recommended standards for patient education, ChatGPT responses averaged a 12th-grade reading level, far exceeding the eighth-grade threshold considered suitable for most U.S. adults [25]. This poses a risk of misinterpretation, particularly among populations with limited health literacy [26]. Prior research has shown that even modest increases in linguistic complexity can impair comprehension and information recall, especially among patients managing chronic conditions [27, 28]. These findings highlight the need to embed readability constraints into AI outputs to ensure accessibility and safety for diverse patient groups.

Regarding quality, ACR fact sheets significantly outperformed ChatGPT outputs in DISCERN scores, indicating higher reliability and depth. Although EQIP scores were comparable, the discrepancy suggests that ChatGPT produces well-structured but often superficial content lacking evidence-based components. This supports previous concerns that ChatGPT's fluency can mask factual inaccuracies or insufficient reasoning [29]. Given the 65% similarity between texts, these differences are not simply due to topic selection but reflect meaningful gaps in content depth, accuracy, and sourcing. Enhancing the clinical credibility of AI tools will require expert validation, transparent sourcing, and regular updates based on current medical guidelines to build trust and minimize misinformation.

Length was another key factor influencing readability. ChatGPT outputs were more than twice as long as ACR materials, contributing to lower readability scores. While length alone does not determine understanding, excessive verbosity can impair focus, elevate cognitive load, and reduce recall – especially among patients with chronic illnesses or limited health literacy [30–32]. Longer texts may also appear thorough while lacking clarity or prioritization. Furthermore, syntactic complexity, technical jargon, and disorganized structure – often present in AI outputs – compound these challenges [33]. Addressing this will require improved summarization algorithms and better integration of user-centered design principles in language model development. Techniques such as reinforcement learning with human feedback or domain-specific fine-tuning could help align AI outputs with medical standards for clarity and conciseness.

Table III. Cosine similarity index values between texts produced by ChatGPT and ACR

Disease	Cosine similarity index
Rheumatoid arthritis	0.74
Systemic lupus erythematosus	0.71
Spondyloarthritis	0.61
Psoriatic arthritis	0.74
Fibromyalgia	0.67
Gout	0.74
Sjögren's disease	0.73
Osteoarthritis	0.74
Scleroderma	0.74
Polymyalgia rheumatica	0.71
Vasculitis	0.61
Inflammatory myopathies	0.69
Reactive arthritis	0.67
Familial Mediterranean fever	0.57
Behçet's disease	0.76

Despite these limitations, ChatGPT shows notable strengths when used in appropriate contexts. It has performed well on standardized medical assessments, demonstrating strong knowledge retrieval and clinical reasoning [34]. ChatGPT has been rated highly empathetic in patient communication, with patients often perceiving its tone as equivalent to that of physicians – even though experts continue to identify shortcomings in accuracy and depth [35]. Its use in clinical documentation has also been promising, particularly for drafting encounter summaries and reducing administrative workload [36]. A recent systematic review found that large language models offer value in tasks such as triage, summarization, and preliminary decision support – when outputs are supervised and constrained by domain knowledge [37]. These findings suggest that ChatGPT, though not suitable for unsupervised patient education, can complement healthcare delivery when embedded in expert-validated workflows.

Our findings also align with recent research in rheumatology showing a divergence between patient and expert evaluations of AI-generated responses. While patients frequently rate ChatGPT's answers as clear and empathetic, clinical experts report gaps in factual accuracy, depth, and nuance [38]. This discrepancy raises concerns about patients' ability to recognize the limitations of AI-generated information, particularly when responses appear polished and fluent [39]. In our study, ACR fact sheets outperformed ChatGPT across multiple quality domains despite moderate textual similarity, indicating that surface-level overlap does not equate to clinical reliability [40]. ChatGPT's lower performance may stem from an inability to prioritize key clinical information, cite sources, or reflect expert judgment. These findings reinforce the importance of treating AI as a complement – not a replacement – for expert-developed educational materials.

As interest in AI-generated health tools grows [41], it remains essential to contextualize their role in healthcare delivery. While tools such as ChatGPT may offer convenient general insights, they are not substitutes for clinical judgment. They lack contextual sensitivity, diagnostic reasoning, and the ability to incorporate patient-specific details such as history, comorbidities, or evolving guidelines – elements critical to safe and effective care [42]. Both ChatGPT and the ACR advise users to consult healthcare professionals for diagnosis and treatment, emphasizing the continued need for clinician oversight in AI-augmented care models. In future, such tools should serve as adjuncts to professional care rather than standalone authorities.

Study limitations

This study has several limitations. First, ChatGPT outputs can vary based on model version, server conditions, and interaction context, all of which were standardized in this study but are subject to variability in real-world use. Second, while validated readability indices (e.g., FKGL, FRE, SMOG) were used, these tools measure only surface-level linguistic complexity and do not capture semantic understanding, cultural relevance, or engagement – factors vital to patient communication [32]. Third, although DISCERN and EQIP provide structured evaluation, both involve subjective scoring, which may introduce evaluator bias [43]. Fourth, the use of a single standardized prompt per condition aimed to enhance internal validity and replicate typical user behavior, though it limited the exploration of prompt variability [40, 44]. Finally, our sample included only 15 English-language RD topics, limiting the generalizability of findings across other specialties, languages, and patient populations. Future studies should assess AI content across prompt variations and model versions and incorporate feedback from clinicians and patients to evaluate clinical relevance, empathy, and trustworthiness in diverse contexts.

Conclusions

This study identified substantial differences in the readability, reliability, and length of ChatGPT-generated texts compared to expert-reviewed materials from the ACR. While ChatGPT outputs exhibited moderate textual similarity to ACR fact sheets, they were significantly more complex, less reliable, and markedly longer, raising concerns about accessibility and trustworthiness in patient education. These findings underscore that, in their current form, large language models are not a substitute for rigorously developed, clinician-reviewed educational content. However, generative AI tools may hold value as adjuncts to healthcare communication when used within well-defined parameters and under professional oversight. Future efforts should focus on improving the factual accuracy, readability, and personalization of AI-generated health information through expert validation pipelines, literacy-aware design constraints, and iterative evaluation frameworks involving both clinicians and patients. This study provides foundational evidence to inform the responsible integration of AI in patient education, particularly in complex chronic disease contexts such as rheumatology.

Acknowledgments

The manuscript preparation process involved using Jenni AI, a generative AI tool, solely to improve grammar, syntax, and clarity without influencing the scientific content, data interpretation, or conclusions. The authors thoroughly reviewed, verified, and edited all AI-generated outputs to ensure accuracy, completeness, and neutrality, as AI-generated content may contain inaccuracies, omissions, or biases. The tool was used under strict human oversight, with full accountability for the final manuscript resting with the authors. The authors take full responsibility for the final content of the manuscript.

Disclosures

Conflict of interest: The authors declare no conflict of interest.

Funding: No external funding.

Ethics approval: As no living creatures were involved, the ethics committee application was not deemed necessary.

Data availability: The data that support the findings of this study are available on request from the corresponding author (Y.E.).

References

1. Kłak A, Raciborski F, Samel-Kowalik P. Social implications of rheumatic diseases. Reumatologia 2016; 54: 73–78, DOI: 10.5114/reum.2016.60216.
2. Huscher D, Merkesdal S, Thiele K, et al. Cost of illness in rheumatoid arthritis, ankylosing spondylitis, psoriatic arthritis and systemic lupus erythematosus in Germany. Ann Rheum Dis 2006; 65: 1175–1183, DOI: 10.1136/ard.2005.046367.
3. Allaire SH, Prashker MJ, Meenan RF. The costs of rheumatoid arthritis. Pharmacoeconomics 1994; 6: 513–522, DOI: 10.2165/00019053-199406060-00005.
4. Schachner T, Keller R, V Wangenheim F. Artificial Intelligence-Based Conversational Agents for Chronic Conditions: Systematic Literature Review. J Med Internet Res 2020; 22: e20701, DOI: 10.2196/20701.
5. Tahri Sqalli M, Al-Thani D. On How Chronic Conditions Affect the Patient-AI Interaction: A Literature Review. Healthcare (Basel) 2020; 8: 313, DOI: 10.3390/healthcare8030313.
6. Altamimi I, Altamimi A, Alhumimidi AS, et al. Artificial Intelligence (AI) Chatbots in Medicine: A Supplement, Not a Substitute. Cureus 2023; 15: e40922, DOI: 10.7759/cureus.40922.
7. Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med 2023; 183: 589–596, DOI: 10.1001/jamainternmed.2023.1838.
8. Fatima A, Shafique MA, Alam K, et al. ChatGPT in medicine: A cross-disciplinary systematic review of ChatGPT's (artificial intelligence) role in research, clinical practice, education, and patient interaction. Medicine (Baltimore) 2024; 103: e39250, DOI: 10.1097/MD.0000000000039250.
9. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel) 2023; 11: 887, DOI: 10.3390/healthcare11060887.
10. Gossec L, Kedra J, Servy H, et al. EULAR points to consider for the use of big data in rheumatic and musculoskeletal diseases. Ann Rheum Dis 2020; 79: 69–76, DOI: 10.1136/annrheumdis-2019-215694.
11. Flood J. ACR presidential address: the wide (and flat) world of rheumatology. Arthritis Rheumatol 2015; 67: 589–594, DOI: 10.1002/art.38997.
12. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med 2007; 4: e296, DOI: 10.1371/journal.pmed.0040296.
13. American College of Rheumatology. Diseases and Conditions. Available at: https://rheumatology.org/patients/diseases-and-conditions (Access: 20.08.2024).
14. Lee H, Hamedi Z, Oliver D, et al. Assessing ChatGPT's potential as a clinical resource for medical oncologists: An evaluation with board-style questions and real-world patient cases. J Clin Oncol 2024; 42 (Suppl 16): e13628-e13628, DOI: 10.1200/JCO.2024.42.16_suppl.e1362.
15. Meskó B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. J Med Internet Res 2023; 25: e50638, DOI: 10.2196/50638.
16. Boles CD, Liu Y, November-Rider D. Readability Levels of Dental Patient Education Brochures. J Dent Hyg 2016; 90: 28–34.
17. Grabeel KL, Russomanno J, Oelschlegel S, et al. Computerized versus hand-scored health literacy tools: a comparison of Simple Measure of Gobbledygook (SMOG) and Flesch-Kincaid in printed patient education materials. J Med Libr Assoc 2018; 106: 38–45, DOI: 10.5195/jmla.2018.262.
18. Weil AG, Bojanowski MW, Jamart J, et al. Evaluation of the quality of information on the Internet available to patients undergoing cervical spine surgery. World Neurosurg 2014; 82: e31–e39, DOI: 10.1016/j.wneu.2012.11.003.
19. Moult B, Franck LS, 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 Expect 2004; 7: 165–175, DOI: 10.1111/j.1369-7625.2004.00273.x.
20. Li B, Han L. Distance Weighted Cosine Similarity Measure for Text Classification. In: Yin H, Tang K, Gao Y, et al. (eds.). Intelligent Data Engineering and Automated Learning – IDEAL 2013. IDEAL 2013. Lecture Notes in Computer Science, vol 8206. Springer-Verlag, Berlin, Heidelberg 2013: 611–618, DOI: 10.1007/978-3-642-41278-3_74.
21. Huang L. Measuring Similarity Between Texts in Python. March 30, 2017. Available at: https://sites.temple.edu/tudsc/2017/03/30/measuring-similarity-between-texts-in-python/ (Access: 10.09.2024).
22. Show word count. Microsoft Support. Available at: https://support.microsoft.com/en-us/office/show-word-count-3c9e6a11-a04d-43b4-977c-563a0e0d5da3#ID0EBBD=Windows (Access: 10.09.2024).

23. Fleckenstein J, Meyer J, Jansen T, et al. Is a Long Essay Always a Good Essay? The Effect of Text Length on Writing Assessment. Front Psychol 2020; 11: 562462, DOI: 10.3389/fpsyg.2020.562462.
24. Le Boedec K. Sensitivity and specificity of normality tests and consequences on reference interval accuracy at small sample size: a computer-simulation study. Vet Clin Pathol 2016; 45: 648–656, DOI: 10.1111/vcp.12390.
25. Handler SJ, Eckhardt SE, Takashima Y, et al. Readability and quality of Wikipedia articles on pelvic floor disorders. Int Urogynecol J 2021; 32: 3249–3258, DOI: 10.1007/s00192-021-04776-0.
26. Daraz L, Morrow AS, Ponce OJ, et al. Readability of Online Health Information: A Meta-Narrative Systematic Review. Am J Med Qual 2018; 33: 487–492, DOI: 10.1177/1062860617751639.
27. Morony S, Webster AC, Buchbinder R, et al. A Linguistic Analysis of Health Literacy Demands of Chronic Kidney Disease Patient Education Materials. Health Lit Res Pract 2018; 2: e1–e14, DOI: 10.3928/24748307-20171227-01.
28. Solnyshkina MI, Harkova EV, Ebzeeva YN. Text content variables as a function of comprehension: Propositional discourse analysis. Russ J Linguist 2023; 27: 938–956, DOI: 10.22363/2687-0088-35915.
29. Tian Tran J, Burghall A, Blydt-Hansen T, et al. Exploring the ability of ChatGPT to create quality patient education resources about kidney transplant. Patient Educ Couns 2024; 129: 108400, DOI: 10.1016/j.pec.2024.108400.
30. Masoni M, Guelfi MR. Going beyond the concept of readability to improve comprehension of patient education materials. Intern Emerg Med 2017; 12: 531–533, DOI: 10.1007/s11739-017-1645-5.
31. Hunt WTN, Sofela J, Mohd Mustapa MF; British Association of Dermatologists' Clinical Standards Unit. Readability assessment of the British Association of Dermatologists' patient information leaflets. Clin Exp Dermatol 2022; 47: 684–691, DOI: 10.1111/ced.15012.
32. Siu AHY, Gibson DP, Chiu C, et al. ChatGPT as a patient education tool in colorectal cancer – an in-depth assessment of efficacy, quality and readability. Colorectal Dis 2025; 27: e17267, DOI: 10.1111/codi.17267.
33. Zeng-Treitler Q, Kim H, Goryachev S, et al. Text characteristics of clinical reports and their implications for the readability of personal health records. Stud Health Technol Inform 2007; 129 (Pt 2): 1117–1121.
34. Yoo JH. Let's Look on the Bright Side of ChatGPT. J Korean Med Sci 2023; 38: e231, DOI: 10.3346/jkms.2023.38.e231.
35. Krusche M, Callhoff J, Knitza J, Ruffer N. Diagnostic accuracy of a large language model in rheumatology: comparison of physician and ChatGPT-4. Rheumatol Int 2024; 44: 303–306, DOI: 10.1007/s00296-023-05464-6.
36. Labinsky H, Nagler LK, Krusche M, et al. Vignette-based comparative analysis of ChatGPT and specialist treatment decisions for rheumatic patients: results of the Rheum2Guide study. Rheumatol Int 2024; 44: 2043–2053, DOI: 10.1007/s00296-024-05675-5. Erratum in: Rheumatol Int 2024; 44: 2055. DOI: 10.1007/s00296-024-05705-2.
37. Yoon J, Cho SK, Choi SR, et al. Expert Consensus on Developing Information and Communication Technology-Based Patient Education Guidelines for Rheumatic Diseases in the Korea. J Korean Med Sci 2025; 40: e67, DOI: 10.3346/jkms.2025.40.e67.
38. Ye C, Zweck E, Ma Z, et al. Doctor Versus Artificial Intelligence: Patient and Physician Evaluation of Large Language Model Responses to Rheumatology Patient Questions in a Cross-Sectional Study. Arthritis Rheumatol 2024; 76: 479–484, DOI: 10.1002/art.42737.
39. Nov O, Singh N, Mann D. Putting ChatGPT's Medical Advice to the (Turing) Test: Survey Study. JMIR Med Educ 2023; 9: e46939, DOI: 10.2196/46939.
40. Guastafierro V, Corbitt DN, Bressan A, et al. Evaluation of ChatGPT's Usefulness and Accuracy in Diagnostic Surgical Pathology. medRxiv 2024.03.12.24304153, DOI: 10.1101/2024.03.12.24304153.
41. Ayo-Ajibola O, Davis RJ, Lin ME, et al. Characterizing the Adoption and Experiences of Users of Artificial Intelligence-Generated Health Information in the United States: Cross-Sectional Questionnaire Study. J Med Internet Res 2024; 26: e55138, DOI: 10.2196/55138.
42. Eysenbach G. Infodemiology: The epidemiology of (mis)information. Am J Med 2002; 113: 763–765, DOI: 10.1016/s0002-9343(02)01473-0.
43. Beheshti M, Toubal IE, Alaboud K, et al. Evaluating the Reliability of ChatGPT for Health-Related Questions: A Systematic Review. Informatics 2025; 12: 9, DOI: 10.3390/informatics12010009.
44. Alapati R, Campbell D, Molin N, et al. Evaluating insomnia queries from an artificial intelligence chatbot for patient education. J Clin Sleep Med 2024; 20: 583–594, DOI: 10.5664/jcsm.10948.