Quality and Comprehensibility of Large Language Model Responses to Common Patient Questions Regarding Musculoskeletal Disorders

Yu Fu, Xi Chen, Mingke You, Lingcheng Wang, Li Wang, Weizhi Liu, Kai Zhou, Gang Chen

Submitted to: Journal of Medical Internet Research
on: June 06, 2024

Table of Contents

Quality and Comprehensibility of Large Language Model Responses to Common Patient Questions Regarding Musculoskeletal Disorders

Yu Fu[1*] MD; Xi Chen[1*] MD, PhD; Mingke You[1] MD; Lingcheng Wang[1] MD; Li Wang[1] MD; Weizhi Liu[1] MD; Kai Zhou[1] MD; Gang Chen[1] MD, PhD

[1]Department of Orthopedics and Orthopedic Research Institute West China Hospital Chengdu CN
[*]these authors contributed equally

Corresponding Author:
Gang Chen MD, PhD
Department of Orthopedics and Orthopedic Research Institute
West China Hospital
West China Hospital Chengdu China
Chengdu
CN

Abstract

Background: Artificial intelligence (AI) and large language models (LLMs) are emerging as the transformative force in various fields, notably in medicine. Their effectiveness in creating physical exercise rehabilitation program and providing information on musculoskeletal (MSK) disorders has yet to be fully explored.

Objective: To assess the quality and readability of an LLM's responses to consultation questions addressing the various phases throughout the entire clinical process experienced by patients with chronic musculoskeletal disorders.

Methods: This cross-sectional study retrieved frequently asked questions from Google (accessed September 3 to October 24, 2023) and randomly selected 25 adult patients suffering from chronic musculoskeletal pain. Three different clinical scenario questions were designed to simulate the entire process of real-world clinical consultations. These questions were used as queries for an AI LLM, ChatGPT version 4.0 (accessed September 23 to December 24, 2023), to prompt LLM-generate responses. The quality of the responses was evaluated by two independent orthopedic clinicians with DISCERN instrument, and the readability was assessed on the WedFX tool website (accessed December 14 to December 20, 2023). Statistical analysis was conducted from January to April 2024.

Results: Of the 98 generated programs, the response format was relatively fixed, based on the queries provided to the LLM. The mean (SD) DISCERN scores assigned by the two doctors were 52.49 (8.57) and 51.50 (8.64), respectively, with all scores ranging from 32 to 67. The analysis of variance between the two physicians (p=0.42) and among the five musculoskeletal disorders (p=0.08) showed no significant difference. The Cohen ? coefficient was calculated to be 0.73 (95% CI ?0.710 to 0.756), signifying good internal agreement, while Cronbach's ? value is 0.834, indicating good reliability. The mean (SD) scores across the six readability tools for all programs were as follows: FRES: 53.04 (16.23), FKGL: 10.15 (3.42), GF: 12.52 (3.41), SMOG: 9.87 (2.75), CLI: 13.28 (1.97), ARI: 10.48 (3.80). All these tools' readability score were significantly different (worse) than the recommended reading level (p<0.05).

Conclusions: In this cross-sectional study, the LLM generated high quality physical exercise prescriptions throughout the entire consultation process of patients with musculoskeletal disorders. However, the lack of supporting materials and the readability of the responses was no sufficiently public-friendly. These findings suggest that, with professional physician evaluation and some improvements, LLMs could be potentially and widely applied in the field of orthopedics in the future.

(JMIR Preprints 06/06/2024:62975)
DOI: https://doi.org/10.2196/preprints.62975

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Quality and Comprehensibility of Large Language Model Responses to Common Patient Questions Regarding Musculoskeletal Disorders

Yu Fu, MD[12&], Xi Chen, MD, PhD[12&], Mingke You, MD[12], Li Wang, MD[12], Weizhi Liu, MD[12], Kai Zhou, MD[12], Lingcheng Wang, MD[12], Gang Chen, MD, PhD[12*]

[1]Sports Medicine Center, West China Hospital, Sichuan University, Chengdu, China.

[2]Department of Orthopedics and Orthopedic Research Institute, West China Hospital, Sichuan University, Chengdu, China.

[&]These authors contributed equally to this work and should be considered co-first authors.

Corresponding Author:

Gang Chen

Sports Medicine Center, West China Hospital, West Chian School of Medicine, Sichuan University, Chengdu, Sichuan, China

Chengdu

CN

Abstract

Background:

Artificial intelligence (AI) and large language models (LLMs) are emerging as the transformative force in various fields, notably in medicine. Their effectiveness in creating physical exercise rehabilitation program and providing information on musculoskeletal (MSK) disorders has yet to be fully explored.

Objective:

To assess the quality and readability of an LLM's responses to consultation questions addressing the various phases throughout the entire clinical process experienced by patients with chronic musculoskeletal disorders.

Methods:

This cross-sectional study retrieved frequently asked questions from Google (accessed September 3 to October 24, 2023) and randomly selected 25 adult patients suffering from chronic musculoskeletal pain. Three different clinical scenario questions were designed to simulate the entire process of real-world clinical consultations. These questions were used as queries for an AI LLM, ChatGPT version 4.0 (accessed September 23 to December 24, 2023), to prompt LLM-generate responses. The quality of the responses was evaluated by two independent orthopedic clinicians with DISCERN instrument, and the readability was assessed on the WedFX tool website (accessed December 14 to December 20, 2023). Statistical analysis was conducted from January to April 2024.

Results:

Of the 98 generated programs, the response format was relatively fixed, based on the queries provided to the LLM. The mean (SD) DISCERN scores assigned by the two doctors were 52.49 (8.57) and 51.50 (8.64), respectively, with all scores ranging from 32 to 67. The analysis of

variance between the two physicians (p=0.42) and among the five musculoskeletal disorders (p=0.08) showed no significant difference. The Cohen κ coefficient was calculated to be 0.73 (95% CI −0.710 to 0.756), signifying good internal agreement, while Cronbach's α value is 0.834, indicating good reliability. The mean (SD) scores across the six readability tools for all programs were as follows: FRES: 53.04 (16.23), FKGL: 10.15 (3.42), GF: 12.52 (3.41), SMOG: 9.87 (2.75), CLI: 13.28 (1.97), ARI: 10.48 (3.80). All these tools' readability score were significantly different (worse) than the recommended reading level (p<0.05).

Conclusions:

In this cross-sectional study, the LLM generated high quality physical exercise prescriptions throughout the entire consultation process of patients with musculoskeletal disorders. However, the lack of supporting materials and the readability of the responses was no sufficiently public-friendly. These findings suggest that, with professional physician evaluation and some improvements, LLMs could be potentially and widely applied in the field of orthopedics in the future.

Keywords: Large Language Model; Musculoskeletal Disorder; Physical Exercise; GPT-4; Rehabilitation Program

Introduction

Musculoskeletal (MSK) disorders present a huge challenge to patients, clinicians, researchers, and governments, ranking as the leading cause of disability worldwide and imposing a significant societal burden[1]. Among individuals suffering from chronic pain, chronic MSK disorders are the most frequent diagnosis, accounting for 70% to 80% of cases[2]. The most common types of chronic MSK pain are chronic low back pain (LBP), shoulder pain (SP), neck

pain (NP), and pain associates with hip and knee osteoarthritis (OA)[1,3]. LBP is a globally recognized health challenge and encompasses various types of pain, including nociceptive, neuropathic, and nociplastic pain, which often coexist. Hip and knee OA are highly prevalent globally, leading to substantial costs and mortality[4]. Shoulder and neck pain exert considerable socioeconomic and personal burdens due to their high disability rates and prevalence. Physical exercise (PE) plays a pivotal role in both the prevention and rehabilitation of various diseases, including ischemic stroke, cardiovascular diseases, pulmonary conditions, diabetes, Parkinson's disease, MSK disorders, and OA[5–10]. Recommended by clinical guidelines for LBP, PE, either alone or combined with educational interventions, plays a crucial role in both primary and secondary prevention and is considered the primary non-pharmacological treatment[11]. In managing pain associated with knee OA, the efficacy of PE is comparable to that of oral analgesics[12] and is widely utilized in pain management. PE has also been proven to be effective in alleviating pain and preventing further deterioration, supported by robust evidence[13,14]. In summary, following established treatment guidelines, PE is acknowledged as an effective intervention for various MSK conditions and typically administered by physical therapists.

As the global population grows and ages, the demand of the PE increased rapidly. However, it cannot be filled due to the shortage of physical therapists and medical resources[15] in many areas of the world especially in many low- and middle-income countries. In contrast, developed nations such as the United States generally provide sufficient PE rehabilitation services[16]. Effective exercise prescriptions may vary due to personally needs and individual characteristics such as ages, contraindications, and body mass index (BMI)[17]. Nevertheless, these factors make it hard to get personalized PE treatment plan for those suffering from MSK disorders. The scarcity of rehabilitation services and personal preferences often drive individuals to seek PE information online rather than consulting a healthcare professional. However, relevant

literature suggests that much of the medical information found on the internet is inaccurate, misleading, and often contains unprofessional advice[18,19]. These factors significantly complicate the process for patients to obtain a personalized, professional, and high-quality PE rehabilitation plan for the recovery from MSK disorders.

Generative artificial intelligence (AI) and large language model (LLM) is emerging as a transformative force in various fields, notably in medicine[20]. ChatGPT-4 (Generative Pre-trained Transformer Version 4), a popular LLM chatbot capable of processing both images and text to generate humanized responses, including text, images and computer code, was unveiled by an AI company Open AI on March 14, 2023[21]. GPT-4 is the most advanced and updated version, demonstrating on average a 60% advantage in accuracy and problem-solving capabilities compared to GPT-3.5[22]. GPT-4 has been applied extensively across various medical scenarios, including medical note-taking, consultations and searches for medical knowledge, based on its wildly training from internet-based sources such as medical texts and research papers[20]. Consequently, GPT-4 has the potential to deliver professional and scientific knowledge on PE along with certain suggestion. Additionally, it could help physical therapists to provide more efficient services by analyzing patient data and generating personalized rehabilitation program as a reference. A study assessed the quality and readability of information related to anterior cruciate ligament injuries indicated that ChatGPT-4 has promising applications in orthopedics[23]. However, there is a notable absence of assessments for patient-centered and scenario-based PE rehabilitation programs developed by GPT-4 within the orthopedic discipline. This study aims to investigate the efficacy of an LLM throughout the entire rehabilitation consulting process for musculoskeletal disorders patients. It examines whether the AI chatbot can deliver accurate and comprehensible knowledge during both pre-diagnosis

and post-diagnosis phases and assesses its capability to generate professional, personalized, and high-quality PE programs.

Methods

This study focus on five most prevalent types of chronic MSK disorders : LBP, SP, NP, hip and knee OA[1,3]. For each disorder, we designed requests based on subtypes defined in relevant guidelines and various clinical scenarios to get more detailed exercise plans. We followed the requests and asked GPT-4 to provide PE rehabilitation programs for quality assessment. In accordance with the guidelines from the American College of Sports Medicine (ACSM), we adopt the FITT-VP principle as a framework for exercise prescriptions, encompassing exercise frequency, intensity, duration, type, overall volume or total intervention length and progression[24,25]. Due to the limitations of GPT-4, we did not include referred pain in our study.

Clinical Scenario Simulation

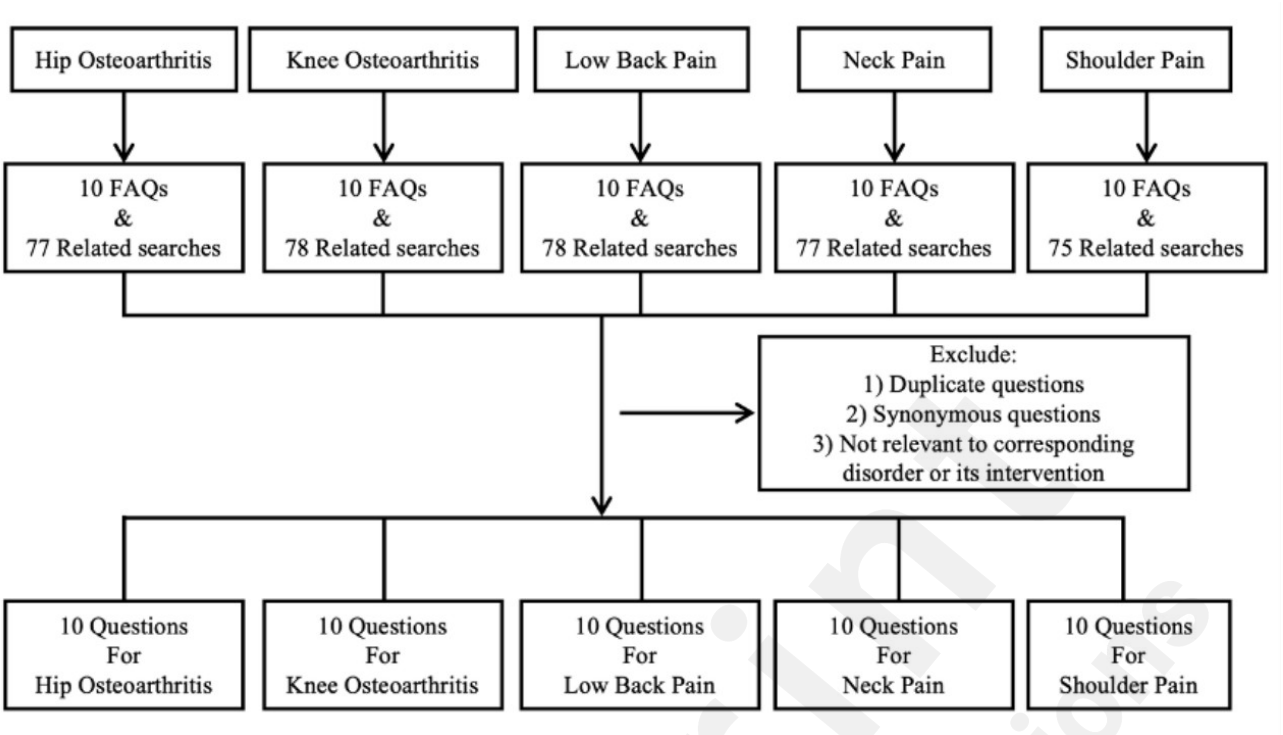
The consulting questions were categorized based on the content, volume and detail of information provided to GPT-4, designed to simulate potential clinical scenarios. We simulated three counseling situations: pre-diagnosis information gathering (Scenario-1), post-diagnosis clinical inquiries (Scenario-2), and post-diagnosis individual case analysis (Scenario-3).

Scenario-1

In the pre-diagnosis searching stage (Scenario-1), we simulated a scene where patients experienced early symptoms such as chronic pain but had not yet been diagnosed with specific disease. During this stage, we utilized the most frequently asked questions (FAQs) associated with the five MSK disorders and categorized by topic and type. To minimize the influence of

personalized search algorithms, we used the new-installed Google Chrome browser (Version 115.0.5790.3). A “new-installed browser” refers to one where all site data, cached images, cookies, files, and browsing history have been cleared. Moreover, after searching for each MSK disorder type, we reinstalled the browser to eliminate any residual effects from previous searches. We extracted the top 10 FAQs by entering disorder-specific terms and augmented these with questions from the "Related Searches" section on Google Web Search, which displays queries closely related to the clicked question. Inclusion criteria for questions were: 1) inclusion any of the terms "low back pain," "neck pain," "shoulder pain," "knee osteoarthritis," or "hip osteoarthritis"; and 2) mention of "therapy," "relief," "exercises," or "stretches." Questions were excluded if they were duplicates, synonymous, or irrelevant to the corresponding disorder or its intervention (e.g., "How long should you lay on a heating pad for back pain?"). An overview of the Scenario-1 question selection strategy is presented in (**Figure 1**). The selected Scenario-1 stage questions provided to GPT-4 are documented in **eTable 1 in Supplement 1**.

Selecting strategy and used the frequently asked questions for pre-diagnosis searching stage (Scenatio-1 stage).



Scenario-2

In the post-diagnosis clinical searching stage (Scenario-2), we simulated a scenario where patients consult physical therapists after being diagnosed with a specific disorder (disease), subtype, or tested with a typical clinical stage. During this process, patients are usually aware of their diagnosis, and physical therapists generally adhere to the FITT-VP principle to devise rehabilitation programs. In this scenario, we provided GPT-4 with the patients' diagnostic outcomes, instructed it to follow the FITT-VP framework to generate physiotherapy plans. Additionally, we asked GPT-4 to enumerate the objectives and rationales for each PE program to facilitate further quality evaluation. The only difference of the consultation queries was the specific diagnosis of the patients. The consultation prompts were structured as follows: “I was diagnosed as XXX, please according to the FITT-VP principle give me a PE rehabilitation plan and list the goal and reason of each stage.” where “XXX” represented the disease or disorder. The queries provided to GPT-4 in Scenario-2 are shown in **eTable 2 in Supplement 1**.

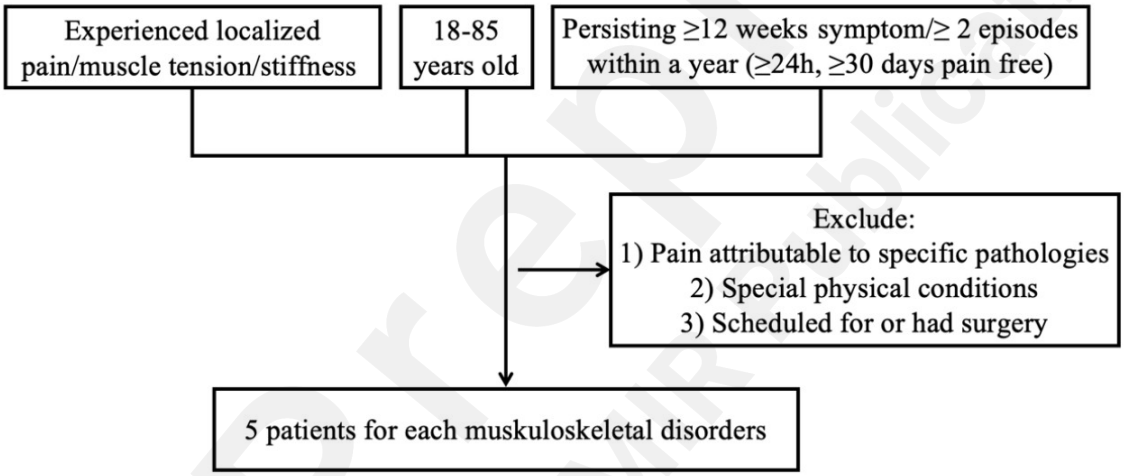
Scenario-3

In the post-diagnosis individual case analysis stage (Scenario-3), we simulated consultations of patients to physical therapists to obtain individualized PE rehabilitation plans according to specific diagnoses. During this process, physical therapists commonly formulate personalized PE prescription tailored to the patients' unique conditions, with patient information systematically recorded and collated in a standardized format. We followed a commonly used clinical format for creating PE rehabilitation program (**eFigure 1 in Supplement 1**), instructing GPT-4 to develop treatment plans based on detailed patient profiles and medical conditions. The directive also specified adherence to the FITT-VP principle, including the reasons and goals for each stage of the rehabilitation program. The consultation prompt was structured as follows: "According to the patient's information, please generate a personalized detailed PE rehabilitation plan and follow the FITT-VP principle. Please list the reason and goal for each phase of the plan. After generating the plan, please check it. Here is the patient information you will use."

In Scenario-3, patients were eligible for inclusion if they 1) experienced localized pain, muscle tension, or stiffness in the corresponding region, with or without referred pain to the legs; 2) were aged between 18 and 85 years; 3) had symptoms persisting for more than 12 weeks or had experienced at least two episodes within a year, each lasting more than 24 hours, with more than 30 days pain-free between. Exclusion criteria included pain attributable to specific pathologies such as fractures, ankylosing spondylitis, spondyloarthritis, infections, neoplasms, or metastasis, or conditions like pregnancy, as well as patients who were scheduled for or had undergone surgery. **Figure 2** presents an overview of the patient selection strategy for Scenario-3. A summary of the patients' information provided to GPT-4 at this stage is displayed

in **eTable 3 in Supplement 1**. Ethical approval from the institutional review board and all patients' written informed consent were acquired.

Inclusion and exclusion criteria for patient selection during the post-diagnosis individual case analysis phase (Scenario-3 stage).



Question Input

During the Scenario-1 and Scenario-2 phases, we directly input structured questions into GPT-4 to generate corresponding treatment plans. In the Scenario-3, due to the incorporation of patient data and the sensitivity of medical information, a meticulous anonymization process is followed before inputting any clinical records into GPT-4. Furthermore, prior to conducting this study, all patient medical reports were finished, and all treatment recommendations or related content generated by GPT-4 are not associated with actual clinical practice. To ensure the

independence of GPT-4's assessment for each question, we input new questions into a new, trackless page after each interaction, free from previous dialogues with GPT-4 and unaffected by any other website influences.

Statistical Analysis

Quality Assessment

To assess the quality of each rehabilitation program generated by GPT-4, two seasoned orthopedic clinicians independently evaluated each program using the DISCERN tool, a validated and reliable tool for judging health information on treatment options[26]. The DISCERN instrument comprises 16 distinct questions (**Table 1**), with each question rated on a 5-point scale (1=definitely no, 5=definitely yes). The questions are organized into three sections: Questions 1 to 8 focus on the reliability of the content, Questions 9 to 15 examine the quality of treatment information, and Question 16 evaluates the overall quality of the treatment plan. The final score of a program is the average of all question scores; a higher score indicates better quality and greater clinical relevance. Online health information on treatment choices is rated by the total DISCERN score, which ranges from 16 to 80, and is categorized as excellent (80 to 63), good (62 to 51), fair (50 to 39), poor (38 to 27), and very poor (26 to 16)[27].

DISCERN Questions

ID	DISCERN question
Q1	Are the aims clear?
Q2	Does it achieve its aims?
Q3	Is it relevant?
Q4	Is it clear what sources of information were used to compile the publication (other than the author or producer)?
Q5	Is it clear when the information used or reported in the publication was produced?
Q6	Is it balanced and unbiased?
Q7	Does it provide details of additional sources of support and information?
Q8	Does it refer to areas of uncertainty?
Q9	Does it describe how each treatment works?

Q10	Does it describe the benefits of each treatment?
Q11	Does it describe the risks of each treatment?
Q12	Does it describe what would happen if no treatment is used?
Q13	Does it describe how the treatment choices affect overall quality of life?
Q14	Is it clear that there may be more than one possible treatment choice?
Q15	Does it provide support for shared decision-making?
Q16	Based on the answers to all of the above questions, rate the overall quality of the publication as a source of information about treatment choices

Interrater Reliability Test

To determine the concordance among evaluators, we calculated the Cohen' kappa (κ) coefficients of the DISCERN scores of each rehabilitation plan to measure interrater reliability, considering the possibility of chance agreement. A Cohen κ value of 1 indicates perfect agreement, whereas a κ value of 0 indicates agreement that's entirely by chance. The agreement levels are categorized as follows: excellent agreement for $\kappa > 0.80$, good agreement for $0.60 < \kappa \leq 0.80$, moderate agreement for $0.40 < \kappa \leq 0.60$, fair agreement for $0.20 < \kappa \leq 0.40$, and poor agreement for $\kappa \leq 0.20$[28].

Internal Consistency Test

To assess the internal consistency of the items, Cronbach's alpha value was calculated along with a 95% confidence interval (CI), based on the scores from 16 DISCERN questions evaluated by two physicians. An alpha value >0.70 indicates good reliability, whereas a value <0.20 signifies poor reliability.

Readability Evaluation

To prevent misinterpretations and confusion often encountered by patients due to obscure medical terminology, it is necessary to evaluate the readability of generated rehabilitation programs and assess the risk of bias. To minimize human error, the responses generated by GPT-4 were input into the online readability test tool, WebFX (https://www.webfx.com/tools/read-able/). WebFX calculates the readability of the text, providing scores according to six readability formulas (**Table 2**): Flesch-Kincaid Grade Level (FKGL), Flesch Reading Ease Score (FRES), Simplified Measure of Gobbledygook (SMOG), Gunning Fog (GF), Automated Readability Index (ARI), and Coleman-Liau Index (CLI). These formulas have been validated for assessing the readability of healthcare and patient information[29]. FKGL and FRES primarily assess average sentence length and syllables per word[30,31]. GF considers sentence length and the number of polysyllabic words which containing three or more syllables[32]. SMOG evaluates polysyllabic words in three 10-sentence samples from the beginning, middle, and end of the text[33]. CLI and ARI are based on the number of letters and average sentence length, incorporating a correction factor for texts containing fewer than 30 sentences[34,35].

Calculating Formula for Each Readability Test Tools

Tool name	Calculating formula
FKGL	$FKGL\ Score = 0.39\left(\frac{total\ words}{total\ sentences}\right) + 11.8\left(\frac{total\ syllables}{total\ words}\right) - 15.59$
FRES	$FRES\ Score = 206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) - 84.6\left(\frac{total\ syllables}{total\ words}\right)$
SMOG	$SMOG\ Score = 1.043\sqrt{\left[\left(total\ polysyllabic\ words\right) \times \left(\frac{30}{total\ sentences}\right)\right] + 3.1291}$
GF	$GF\ Score = 0.4\left[\left(\frac{total\ words}{total\ sentences}\right) + 100\left(\frac{total\ polysyllabic\ words}{total\ words}\right)\right]$
ARI	$ARI\ Score = 4.71\left(\frac{total\ characters}{total\ words}\right) + 0.5\left(\frac{total\ words}{total\ sentences}\right) - 21.43$
CLI	$CLI\ Score = 5.89\left(\frac{total\ characters}{total\ words}\right) - 0.3\left(\frac{total\ sentences}{total\ words}\right) - 15.8$

In accordance with guidelines from the US Department of Health and Human Services (USDHHS), healthcare information should be written at a level comprehensible to 11–12-year-olds, roughly equivalent to the sixth grade. In this study, materials considered easy to read achieved a minimum score of 80 on FRES[36]. And for FKGL, SMOG, GF, CLI and ARI, the same level was capped at a score of 6.9 or lower. The USDHHS categorizes reading levels into three groups: easy (below 12 years old), average (12 to 15 years old), and difficult (above 15 years old). The scale for each readability metric used in this study are detailed in **Table 3**.

Scale for Six Readability Test Tools

FRES readability scale score	FKGL, SMOF, CLI, GF or ARI readability scale score	Age	USDHHS reading level
90-100	0-1	6-7 yr old	Very Easy
	2-3	7-8 yr old	
	3-4	8-9 yr old	
	4-5	9-10 yr old	
	5-6	10-11 yr old	
80-89*	6-7*	11-12 yr old*	Easy*
70-79	7-8	12-13 yr old	Fairly Easy
60-69	8-9	13-14 yr old	Average
	9-10	14-15 yr old	
50-59	10-11	15-16 yr old	Fairly Difficult
	11-12	16-17 yr old	
	12-13	17-18 yr old	
30-49	13-17	18-22 yr old	Difficult
0-29	17+	22+yr old	Very difficult

*Recommended reading level

Statistical Analysis

All statistical analyses were conducted using the Meta package in R (R software version 3.6.0, R Foundation for Statistical Computing Platform). Means and standard errors (SE) were calculated. Differences between groups were evaluated using the two-tailed Welch's t-test, with P values below 0.05 deemed statistically significant.

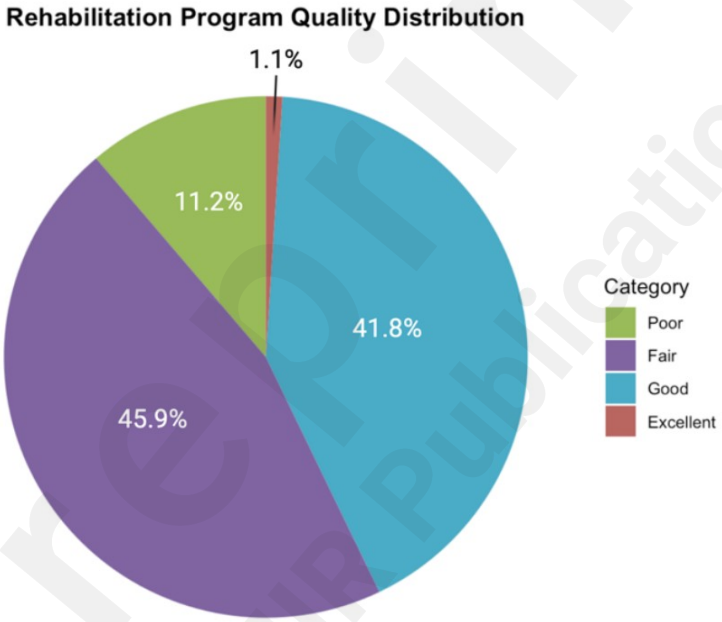
Results

After receiving a rehabilitation query, GPT-4 produced a MSK disorder-related PE plan, with varying formats across each stage. Of the 98 generated results, 51.02% (50/98) are in Scenario-1, 23.47% (23/98) pertain to Scenario-2 and 25.51% (25/98) belong to Scenario-3. The Scenario-1 rehabilitation program is structured into three segments: a basic introduction or instructions, detailed descriptions of exercises and notes on safety precautions. The Scenario-2 plan also consists of three parts: an introduction that explains the FITT-VP principle and its application to PE rehabilitation for the disorder, a detailed framework of the rehabilitation plan and a concluding section with safety notes. In this stage, the detailed framework includes multiple phases outlining the goals, frequency, intensity, time, type, volume, and progression in accordance with the FITT-VP principle. The Scenario-3 PE programs are more comprehensive, encompassing five parts: a title, an explanation of the FITT-VP principle, a multi-phase exercise plan like the structure in Scenario-2, general tips or recommendations and a review section to ensure accuracy and completeness.

All 98 rehabilitation exercise plans were assessed by two independent physicians using the DISCERN instrument. The mean (SD) DISCERN scores assigned by the two doctors were 52.49 (8.57) and 51.50 (8.64), respectively, with all scores ranging from 32 to 67. After calculating the mean score of each program, the finally mean (SD) score of all 98 plans were 51.99 (8.51). Categorized according to DISCERN criteria, the distribution of scores shows as follows: Poor (11/98), Good (45/98), Fair (41/98) and Excellent (1/98), as depicted in the **Figure 3**. 88.8% of the PE rehabilitation programs showed relatively high quality. The analysis of variance (ANOVA) revealed no statistically significant difference between the scores of the two doctors

(p=0.42). However, in the ANOVA and subsequent Tukey's test grouped by scenario, all p-values were below 0.05, indicating there was a statistically significant difference in scores across the scenarios: Scenario-1 scored significantly lower than Scenario-2 and 3, and Scenario-2 was also lower than Scenario-3. Meanwhile, no significant differences were found in the average DISCERN scores across the five types of MSK disorders (p=0.08).

Distribution of the rehabilitaion programs quality according to the DISCERN criteria.



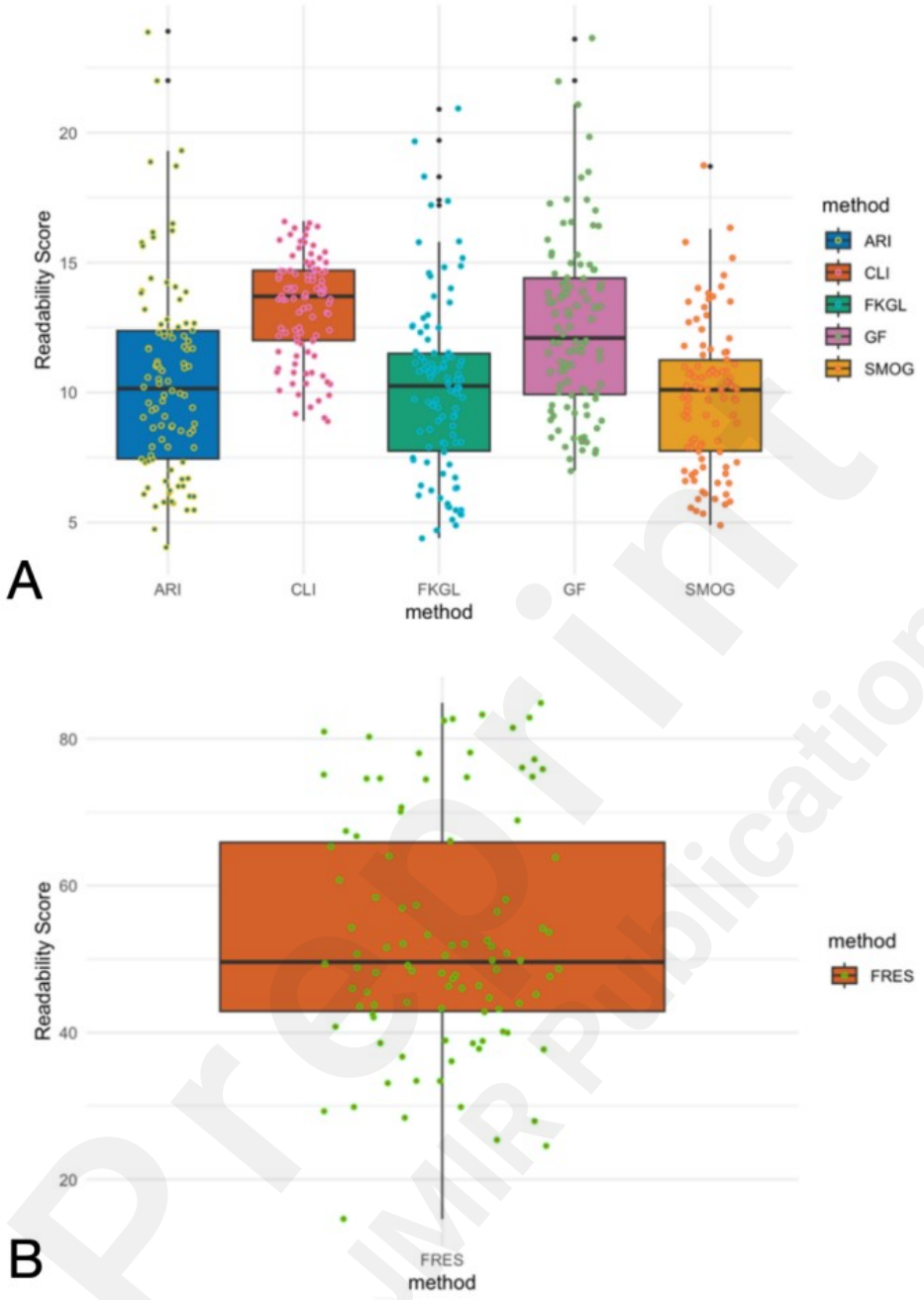
Following the evaluation scores provided by two doctors for the exercise plans, the Cohen κ coefficient was calculated to be 0.73 (95% CI 0.710-0.756), signifying good internal agreement as according to Fleiss[28]. This result demonstrates that the scores derived from the DISCERN tool are reliable between the two independent raters.

When calculating Cronbach's α value, items Q4 and Q5 demonstrated a negative correlation with the principal component, while Q12 exhibited no variance. After excluding Q4, Q5, and

Q12, the final Cronbach's α value is 0.834, indicating good reliability. The 95% CI for the alpha value is rendered meaningless due to the low variance (high constancy) of certain items, resulting in a null SE.

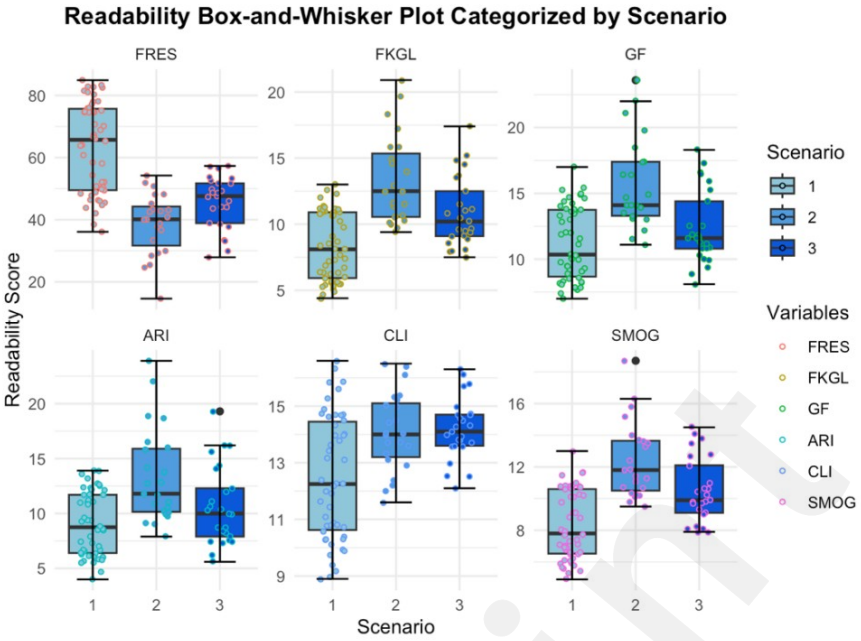
All 98 rehabilitation programs underwent readability testing. The mean (SE) scores across the six readability tools for all programs were as follows: FRES: 53.04 (16.23), FKGL: 10.15 (3.42), GF: 12.52 (3.41), SMOG: 9.87 (2.75), CLI: 13.28 (1.97), ARI: 10.48 (3.80). The distribution of readability scores is illustrated in box-and-whisker plots (**Figure 4**). All six readability assessment tools reported statistically significant results, with p-values less than the standard alpha value of 0.05, indicating that the readability of the three-scenario programs is significantly different (worse) from the “easy to read material” standards described in the previous “Method” section. Subgroup analysis sorted the readability scores by “Scenario” or “Disease.” ANOVA analysis revealed significant differences: the readability of Scenario-1 programs was superior to that of Scenario-2 and 3 in terms of FRES, FKGL, SMOG, and CLI, whereas the readability of Scenario-2 programs was inferior to Scenario-1 and 3 for FKGL, GF, SMOG, and ARI. The distribution of readability scores grouped by Scenario is shown in box-and-whisker plots (**Figure 5**). Despite the variations across the six testing methods, these differences were not disease-specific ($p>0.05$). After proportionally adjusting the FRES score into 0 to 25, the average scores for each disease type and scenario are displayed in a heat plot (**Figure 6**).

Readability Score of all 98 Rehabilitation Programs of Six Formulas

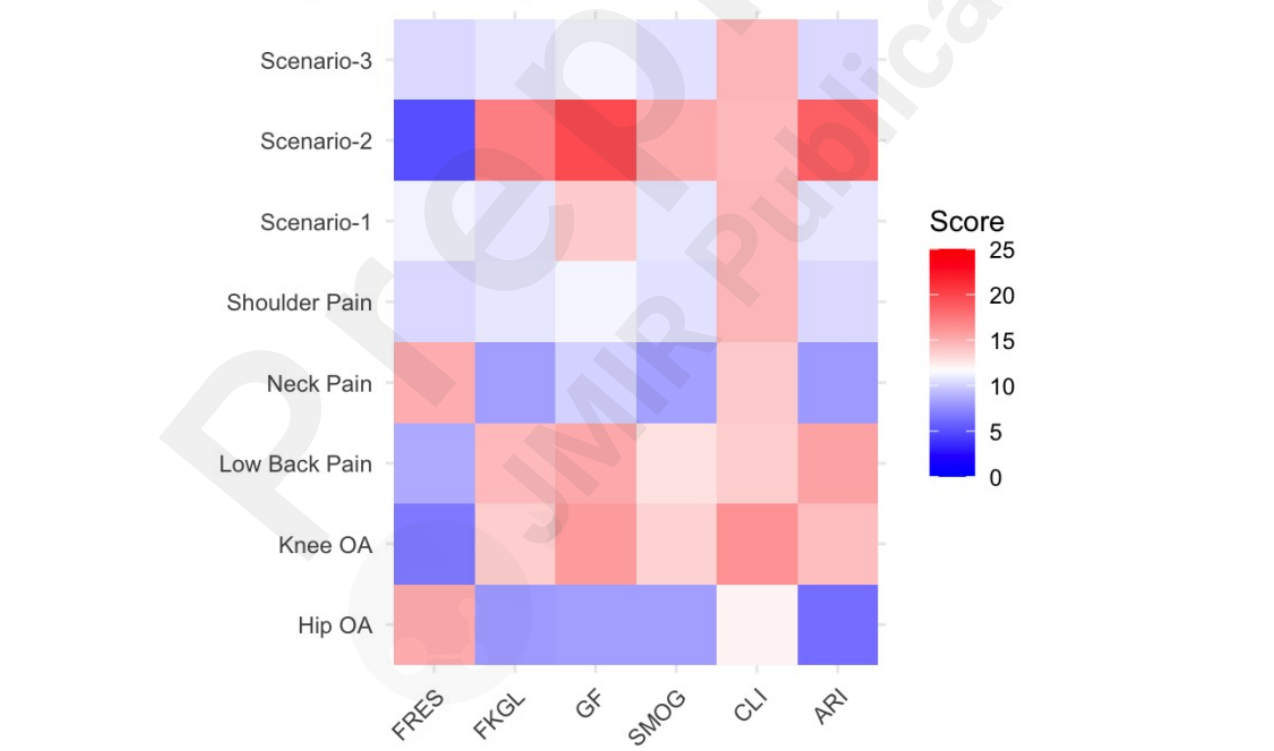


Readability score of all 98 rehabilitation programs of six formulas. (A) Box-and-whisker Plot for ARI, CLI, FKGL, GF and SMOG. (B) Box-and-whisker Plot for FRES.

Readability Box-and-whisker Plot Grouped by Level



Heat Plot of Average Readability Score for Each Disease Type and Scenario



Readability scores for each disease tyoe and Scenario, the white color is assigned value as the mean readablity score of all 98 rehabilitation programs, the higher score indicates the poorer readability.

Discussion

To our knowledge, this is the first study to test the quality and readability of GPT-4's generation on PE rehabilitation program of MSK disorders. Our findings simulated three possible situations that patients who suffered five types of MSK disorders may experience, ultimately got 98 rehabilitation programs after giving the instructions to GPT-4. DISCERN tool was used to evaluate the quality of the generated plans and six readability tools tested the readability. Two professional clinical doctors independently participated in scoring the quality of the programs, the internal consistency of the evaluating tool and the internal reliability of two doctors are both good. Our study demonstrates that whether providing general question or specific patients' information, GPT-4 gave good quality PE programs for all chronic pain stage but relatively difficult to read. Moreover, GPT-4 can generate personalized PE prescriptions based on the specific information of patients with MSK disorders. However, GPT-4 didn't provide the source of the supporting information and lack of the description of what would happen if no treatment were used. All this evidence suggests that GPT-4 can serve as an excellent tool for initial personalized consultations for patients at any stage of musculoskeletal chronic pain and offering high-quality references for clinical doctors.

Over the past decade, the number of internet users and the visits to health-related websites have steadily increased, making the internet a critical source of health information for the public[37]. However, internet health information has numerous issues such as inaccuracies, insufficient detail, a lack of personalization, misleading, and even occurrences of errors. Several studies that utilized DISCERN as an assessment tool have indicated that the quality of health information provided by internet sites is relatively poor[38–40]. This highlights the ongoing lack of an online platform capable of offering accurate, personalized, and high-quality health

information. The AI chatbot GPT-4 has demonstrated potential for use in clinical consultation scenarios and addresses the aforementioned issue; however, further testing is required to assess the quality and readability of its outputs. Consequently, we employed six readability testing tools and the DISCERN instrument to evaluate the responses we received.

In our tests, when requiring outlining FITT-VP principles, setting objectives, and explaining rationales, GPT-4 demonstrated commendable compliance by accurately following our instructions and independently listing the relevant sections. Regardless of the type of MSK disorder or the scenario, the physical therapy rehabilitation plans generated by GPT-4 were consistently effective, without significant errors. This suggests that GPT-4's knowledge base across the five disorder domains does not exhibit significant heterogeneity. Additionally, we observed that as the scenario changed, the quality of the generated results improved, likely due to the varying requirements provided to GPT-4 at each scenario. Scenario-2, compared to Scenario-1, included two additional requirements: adherence to the FITT-VP principles and explain the goals and rationale of the plan. Scenario-3 built upon this by adding anonymized patient information, which provided GPT-4 with more details, significantly enhancing the quality of the generated rehabilitation programs. Additionally, in Scenario-3, GPT-4 considered the chief medical problem, exercise preferences, available equipment, exercise environment, and other personal patient information to generate comprehensive, individualized, and high-quality PE prescriptions. This indicates that the format, content, and quality of GPT-4's output depend on the given instructions. Without specific directives, GPT-4 struggles to provide additional information on that topic. The outputs at Scenario-1 and Scenario-2 align well with our intended purposes, offering relatively high readability, especially in response to the simpler questions posed at Scenario-1, where GPT-4 provided concise answers. The setup at Scenario-3 met our expectations as well, with higher quality but lower readability. During this phase, we

observed an abundance of medical terminology within the treatment plans, which may cause reading challenges for those without professional medical knowledge, yet poses no obstacle for clinicians. This demonstrates that GPT-4 is indeed capable of tailoring its responses based on the object of the conversation, underscoring its potential application in helping patients to access health information and providing recommendations to doctors.

It should be noted that, in the DISERN evaluation process, questions Q4, Q5, and Q12 scored particularly low. Q4 and Q5 pertain to the assessment of information sources for the generated results, while Q12 queries the description of the consequences of lacking a treatment. A reliable information source is important for a high-quality rehabilitation program[41]. Additionally, during the selection process for treatment options, if no alternatives are available, it is vital for patients to understand the potential outcomes clearly. This understanding plays a significant role in safeguarding patients' health and maintaining a positive doctor-patient relationship[42,43]. GPT-4 performed poorly in the above two aspects, likely due to the absence of direct instructions. However, we also observed that for the third results of Scenario-1 regarding Knee OA, under the Q4 scoring item, both doctors awarded a score of 5. Our analysis of GPT-4's responses revealed that although it did not provide an exercise video for Knee OA as requested, it did offer two websites as sources of information. This further corroborates the characteristic of GPT-4 discussed previously, where GPT-4's outputs are highly contingent upon the instructions received. This insight prompts us to take a more comprehensive approach when providing commands to GPT-4, considering all facets of a rehabilitation treatment plan, including requiring it to furnish supporting information and diverse treatment options, thereby enabling GPT-4 to generate high-quality results. However, we recognized that the general public's lack of medical knowledge can complicate their ability to provide comprehensive and precise instructions to GPT-4, thereby impacting the quality of the generated results and posing

potential risks. This represents a significant challenge for GPT-4 in delivering medical services and information to a broader application for public and acts as a barrier to its widespread adoption in healthcare. Encouragingly, Microsoft's recent release of its AI-driven Bing search and the new version of GPT-4, both powered by GPT-4 and capable of including links in responses, may address this issue[44].

In readability test, the readability of the rehabilitation programs generated by GPT-4 was deemed fairly difficult, consistent with findings from earlier studies[45]. Compared to a survey in another study[46], the FRES for results generated by GPT-4 in this study was lower than those from Google, while the FKGL was higher, suggesting that without specific instructions, the readability of treatment plans produced by GPT-4 is inferior to Google. Studies have shown that GPT-4 can simplify health information and make it more readable[47]. Unlike static websites, GPT-4 can tailor simplified content to match the user's comprehension level. However, this simplification process may lead to misunderstandings[48]. Medical terminology is obscure for non-medical professionals, indiscriminate simplification of medical text may lead to further misunderstandings or even alter its original meaning and confused patients[49]. Several studies showed that GPT-4 may fabricate data or even provide incorrect information, which is unacceptable in clinical practice[50]. This indicates the indispensable role of experienced physicians in the delivery of medical services, emphasizing the irreplaceable position of orthopedic doctors in devising, explaining, and implementing PE rehabilitation plans to ensure comprehensive understanding and sensible decision.

Limitations

This study possesses several limitations. First, we enlisted two experienced professional doctors as expert raters during the scoring process. To more robustly validate the quality of therapeutic plans generated by GPT-4, it is essential to involve a larger number of medical experts to minimize the impact of subjective judgments. Our structured questions may have led to a homogeneity in the generated outcomes and making the performance of GPT-4 appear monotonous. Future research could incorporate a broader range of question formats to enrich GPT-4's responses, better simulating the real-world inquiries patients encounter during their diagnosis and treatment process. In Scenario-2, the simulated scenarios might not fully reflect the challenges patients face in actual conditions, as our research positioned doctors as intermediaries, without direct questioning from patients to GPT-4. Another limitation is encountered at Scenario-3, due to the relatively small size of the patient cohort. While this may be deemed acceptable for the initial exploratory phase, more extensive studies are imperative to ensure the findings' applicability on a wider scale. We created a new traceless page and restarted the conversation before providing the questions, this could potentially compromise the completeness of patient history collection by GPT-4, thus affecting its output. Additionally, the AI advanced rapidly, the release of new versions of GPT-4 after our data collection might lead to different outcomes. Despite these challenges, this study provides valuable insights into GPT-4's potential capability in generating PE rehabilitation treatment plans of MSK disorders and offering health-related information. GPT-4 holds promise as an auxiliary consultation tool in orthopedics and might offer valuable suggestions during the creation of PE training plans by clinicians. Although GPT-4 demonstrated commendable performance in this study, it is imperative to emphasize that it cannot replicate the extensive experience and judgment of clinical doctors and necessitating cautious use.

Conclusion

With the advancement of artificial intelligence and large language model, GPT-4, an AI chatbot, has shown the potential application in the medical field. In this study, we evaluated the quality and readability of the physical exercise rehabilitation program created by GPT-4 across three simulated scenarios. The results demonstrated consistently high-quality outputs regardless of the stage of the patient's musculoskeletal disorder, though some readability challenges were noted. Our findings suggest that GPT-4 can comply with specific instructions, demonstrating its capacity to generate appropriate responses as requested. However, without explicit directives, GPT-4 does not offer additional information, highlighting the need for a certain degree of expertise and comprehensive consideration to fully utilize it. As AI LLM continues to improve with updates, it holds the promise to address various issues present in internet health information, serving as a pathway for patients to access high-quality medical information and potentially offering suggestions to clinical doctors when giving the recovery plans. Enhancing the accuracy and professionalism of AI LLM is crucial for its broader application in the future. It is important to note that artificial intelligence cannot replace clinical doctors who have rich experience and professional knowledge. Professional oversight and evaluation by medical professionals are essential during its use to better cater to diverse using.

Abbreviations

LLM: Large Language Model

OA: Osteoarthritis

LBP: Low Back Pain

SP: Shoulder Pain

NP: Neck Pain

MSK: Musculoskeletal

PE: Physical Exercise

GPT-4: Generative Pre-trained Transformer Version 4

BMI: Body Mass Index

AI: Artificial Intelligence

ACSM: American College of Sports Medicine

FAQs: Frequently Asked Questions

CI: Compound Interest

FKGL: Flesch-Kincaid Grade Level

FRES: Flesch Reading Ease Score

SMOG: Simplified Measure of Gobbledegook

GF: Gunning Fog

ARI: Automated Readability Index

CLI: Coleman-Liau Index

USDHHS: US Department of Health and Human Services

SE: Standard Errors

ANOVA: Analysis of Variance

References

1. Vos T, Abajobir AA, Abate KH, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*.

2017;390(10100):1211-1259. doi:10.1016/S0140-6736(17)32154-2

2. Institute of Medicine (US) Committee on Advancing Pain Research, Care, andEducation. *Relieving Pain in America: A Blueprint for Transforming Prevention, Care, Education, and Research*. National Academies Press (US); 2011. Accessed July 8, 2023. http://www.ncbi.nlm.nih.gov/books/NBK91497/

3. Smith E, Hoy DG, Cross M, et al. The global burden of other musculoskeletal disorders: estimates from the Global Burden of Disease 2010 study. *Ann Rheum Dis*. 2014;73(8):1462-1469. doi:10.1136/annrheumdis-2013-204680

4. Katz JN, Arant KR, Loeser RF. Diagnosis and Treatment of Hip and Knee Osteoarthritis: A Review. *JAMA*. 2021;325(6):568. doi:10.1001/jama.2020.22171

5. Kyu HH, Bachman VF, Alexander LT, et al. Physical activity and risk of breast cancer, colon cancer, diabetes, ischemic heart disease, and ischemic stroke events: systematic review and dose-response meta-analysis for the Global Burden of Disease Study 2013. *The BMJ*. 2016;354:i3857. doi:10.1136/bmj.i3857

6. Lear SA, Hu W, Rangarajan S, et al. The effect of physical activity on mortality and cardiovascular disease in 130 000 people from 17 high-income, middle-income, and low-income countries: the PURE study. *The Lancet*. 2017;390(10113):2643-2654. doi:10.1016/S0140-6736(17)31634-3

7. McLoughlin RF, Clark VL, Urroz PD, Gibson PG, McDonald VM. Increasing physical activity in severe asthma: a systematic review and meta-analysis. *Eur Respir J*. 2022;60(6):2200546. doi:10.1183/13993003.00546-2022

8. Ascherio A, Schwarzschild MA. The epidemiology of Parkinson's disease: risk factors and

prevention. *Lancet Neurol*. 2016;15(12):1257-1272. doi:10.1016/S1474-4422(16)30230-7

9. Maestroni L, Read P, Bishop C, et al. The Benefits of Strength Training on Musculoskeletal System Health: Practical Applications for Interdisciplinary Care. *Sports Med*. 2020;50(8):1431-1450. doi:10.1007/s40279-020-01309-5

10. Skou ST, Roos EM. Physical therapy for patients with knee and hip osteoarthritis: supervised, active treatment is current best practice. *Clin Exp Rheumatol*. Published online 2019.

11. Knezevic NN, Candido KD, Vlaeyen JWS, Van Zundert J, Cohen SP. Low back pain. *The Lancet*. 2021;398(10294):78-92. doi:10.1016/S0140-6736(21)00733-9

12. Henriksen M, Hansen JB, Klokker L, Bliddal H, Christensen R. Comparable effects of exercise and analgesics for pain secondary to knee osteoarthritis: a meta-analysis of trials included in Cochrane systematic reviews. *J Comp Eff Res*. 2016;5(4):417-431. doi:10.2217/cer-2016-0007

13. Cohen SP, Hooten WM. Advances in the diagnosis and management of neck pain. *BMJ*. Published online August 14, 2017:j3221. doi:10.1136/bmj.j3221

14. Steuri R, Sattelmayer M, Elsig S, et al. Effectiveness of conservative interventions including exercise, manual therapy and medical management in adults with shoulder impingement: a systematic review and meta-analysis of RCTs. *Br J Sports Med*. 2017;51(18):1340-1347. doi:10.1136/bjsports-2016-096515

15. Chen X, Giles J, Yao Y, et al. The path to healthy ageing in China: a Peking University–Lancet Commission. *The Lancet*. 2022;400(10367):1967-2006. doi:10.1016/S0140-6736(22)01546-X

16. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*. 2020;396(10267):2006-2017. doi:10.1016/S0140-6736(20)32340-0

17. Goh SL, Persson MSM, Stocks J, et al. Efficacy and potential determinants of exercise therapy in knee and hip osteoarthritis: A systematic review and meta-analysis. *Ann Phys Rehabil Med*. 2019;62(5):356-365. doi:10.1016/j.rehab.2019.04.006

18. Tang H, Ng JHK. Googling for a diagnosis—use of Google as a diagnostic aid: internet based study. *BMJ*. 2006;333(7579):1143-1145. doi:10.1136/bmj.39003.640567.AE

19. Jalees R. Accuracy of Medical Information on the Internet - Scientific American Blog Network. Published August 2, 2012. Accessed July 27, 2023. https://blogs.scientificamerican.com/guest-blog/accuracy-of-medical-information-on-the-internet/

20. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N Engl J Med*. Published online 2023.

21. Sanderson K. GPT-4 is here: what scientists think. *Nature*. 2023;615(7954):773-773. doi:10.1038/d41586-023-00816-5

22. Currie GM. GPT-4 in Nuclear Medicine Education: Does It Outperform GPT-3.5? *J Nucl Med Technol*. 2023;51(4):314-317. doi:10.2967/jnmt.123.266485

23. Kaarre J, Feldt R, Keeling LE, et al. Exploring the potential of ChatGPT as a supplementary tool for providing orthopaedic information. *Knee Surg Sports Traumatol Arthrosc*. 2023;31(11):5190-5198. doi:10.1007/s00167-023-07529-2

24. Piercy KL, Troiano RP, Ballard RM, et al. The Physical Activity Guidelines for Americans. *JAMA*. 2018;320(19):2020. doi:10.1001/jama.2018.14854

25. Piercy KL, Troiano RP. Physical Activity Guidelines for Americans From the US Department of Health and Human Services: Cardiovascular Benefits and Recommendations. *Circ Cardiovasc Qual Outcomes*. 2018;11(11):e005263. doi:10.1161/CIRCOUTCOMES.118.005263

26. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *J Epidemiol Community Health*. 1999;53(2):105-111. Accessed September 5, 2023. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1756830/

27. Sun F, Yang F, Zheng S. Evaluation of the Liver Disease Information in Baidu Encyclopedia and Wikipedia: Longitudinal Study. *J Med Internet Res*. 2021;23(1):e17680. doi:10.2196/17680

28. Fleiss JL. *Statistical Methods for Rates and Proportions. Second Edition*. Wiley, John and Sons, Incorporated, New York, N.Y.; 1981.

29. Barnett T, Hoang H, Furlan A. An analysis of the readability characteristics of oral health information literature available to the public in Tasmania, Australia. *BMC Oral Health*. 2016;16(1):35. doi:10.1186/s12903-016-0196-x

30. Flesch R. A new readability yardstick. *J Appl Psychol*. 1948;32(3):221-233. doi:10.1037/h0057532

31. Kincaid JP, Fishburne Jr, Robert P. R, Richard L. C, Brad S. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for*

Navy Enlisted Personnel: Defense Technical Information Center; 1975. doi:10.21236/ADA006655

32. Gunning R. The technique of clear writing. *No Title*. Accessed September 9, 2023. https://cir.nii.ac.jp/crid/1130000796418620544

33. Mc Laughlin: SMOG grading-a new readability formula - Google Scholar. Accessed September 9, 2023. https://scholar.google.com/scholar_lookup?journal=J+Read&title=SMOG+grading+-+a+new+readability+formula&volume=12&publication_year=1969&pages=638-646&

34. Coleman: A computer readability formula designed... - Google Scholar. Accessed September 9, 2023. https://scholar.google.com/scholar_lookup?journal=J+Appl+Psychol&title=A+computer+readability+formula+designed+for+machine+scoring&volume=60&publication_year=1975&pages=283-284&

35. Derivation and Validation of the Automated Readability Index for Use with Technical Materials - Edgar A. Smith, J. Peter Kincaid, 1970. Accessed September 9, 2023. https://journals.sagepub.com/doi/abs/10.1177/001872087001200505

36. Edmunds MR, Barry RJ, Denniston AK. Readability Assessment of Online Ophthalmic Patient Information. *JAMA Ophthalmol*. 2013;131(12):1610. doi:10.1001/jamaophthalmol.2013.5521

37. Haluza D, Naszay M, Stockinger A, Jungwirth D. Digital Natives Versus Digital Immigrants: Influence of Online Health Information Seeking on the Doctor–Patient Relationship. *Health Commun*. 2017;32(11):1342-1349. doi:10.1080/10410236.2016.1220044

38. Cassidy JT, Fitzgerald E, Cassidy ES, et al. YouTube provides poor information regarding

anterior cruciate ligament injury and reconstruction. *Knee Surg Sports Traumatol Arthrosc*. 2018;26(3):840-845. doi:10.1007/s00167-017-4514-x

39. Hirasawa R, Saito K, Yachi Y, et al. Quality of Internet information related to the Mediterranean diet. *Public Health Nutr*. 2012;15(5):885-893. doi:10.1017/S1368980011002345

40. Hargrave DR, Hargrave UA, Bouffet E. Quality of health information on the Internet in pediatric neuro-oncology. Published online 2006.

41. Labrecque MS, Ruckdeschel' JC. INFORMATION AND DECISION-MAKING PREFERENCES OF HOSPITALIZED ADULT CANCER PATIENTS.

42. Greenfield S. Expanding Patient Involvement in Care: Effects on Patient Outcomes. *Ann Intern Med*. 1985;102(4):520. doi:10.7326/0003-4819-102-4-520

43. Kaba R, Sooriakumaran P. The evolution of the doctor-patient relationship. *Int J Surg*. 2007;5(1):57-65. doi:10.1016/j.ijsu.2006.01.005

44. Yanagita Y, Yokokawa D, Uchida S, Tawara J. Accuracy of ChatGPT on Medical Questions in the National Medical Licensing Examination in Japan: Evaluation Study. *JMIR Form Res*.

45. Haver HL, Lin CT, Sirajuddin A, Yi PH, Jeudy J. Use of ChatGPT, GPT-4, and Bard to Improve Readability of ChatGPT's Answers to Common Questions About Lung Cancer and Lung Cancer Screening. *AJR Am J Roentgenol*. 2023;221(5):701-704. doi:10.2214/AJR.23.29622

46. Mastrokostas PG, Mastrokostas LE, Emara AK, et al. GPT-4 as a Source of Patient Information for Anterior Cervical Discectomy and Fusion: A Comparative Analysis Against Google Web Search. *Glob Spine J*. Published online March 21, 2024:21925682241241241.

doi:10.1177/21925682241241241

47. Doshi R, Amin KS, Khosla P, Bajaj S, Chheang S, Forman HP. Quantitative Evaluation of Large Language Models to Streamline Radiology Report Impressions: A Multimodal Retrospective Analysis. *Radiology*. 2024;310(3):e231593. doi:10.1148/radiol.231593

48. Monteith S, Glenn T, Geddes JR, Whybrow PC, Achtyes E, Bauer M. Artificial intelligence and increasing misinformation. *Br J Psychiatry*. 2024;224(2):33-35. doi:10.1192/bjp.2023.136

49. Boyle CM. Difference Between Patients' and Doctors' Interpretation of Some Common Medical Terms. *Br Med J*. 1970;2(5704):286-289. Accessed April 12, 2024. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1700443/

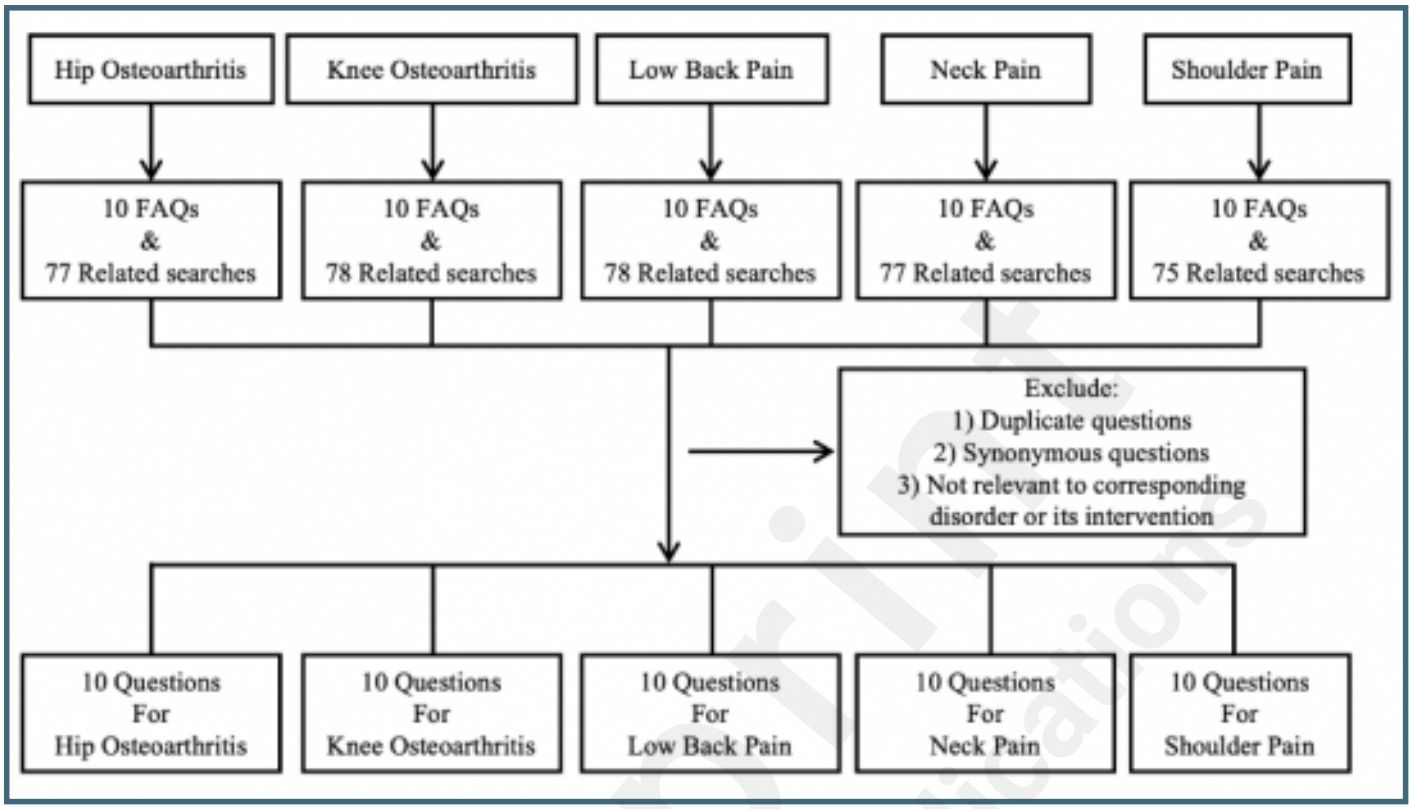
50. Naddaf M. ChatGPT generates fake data set to support scientific hypothesis. *Nature*. 2023;623(7989):895-896. doi:10.1038/d41586-023-03635-w

Supplementary Files

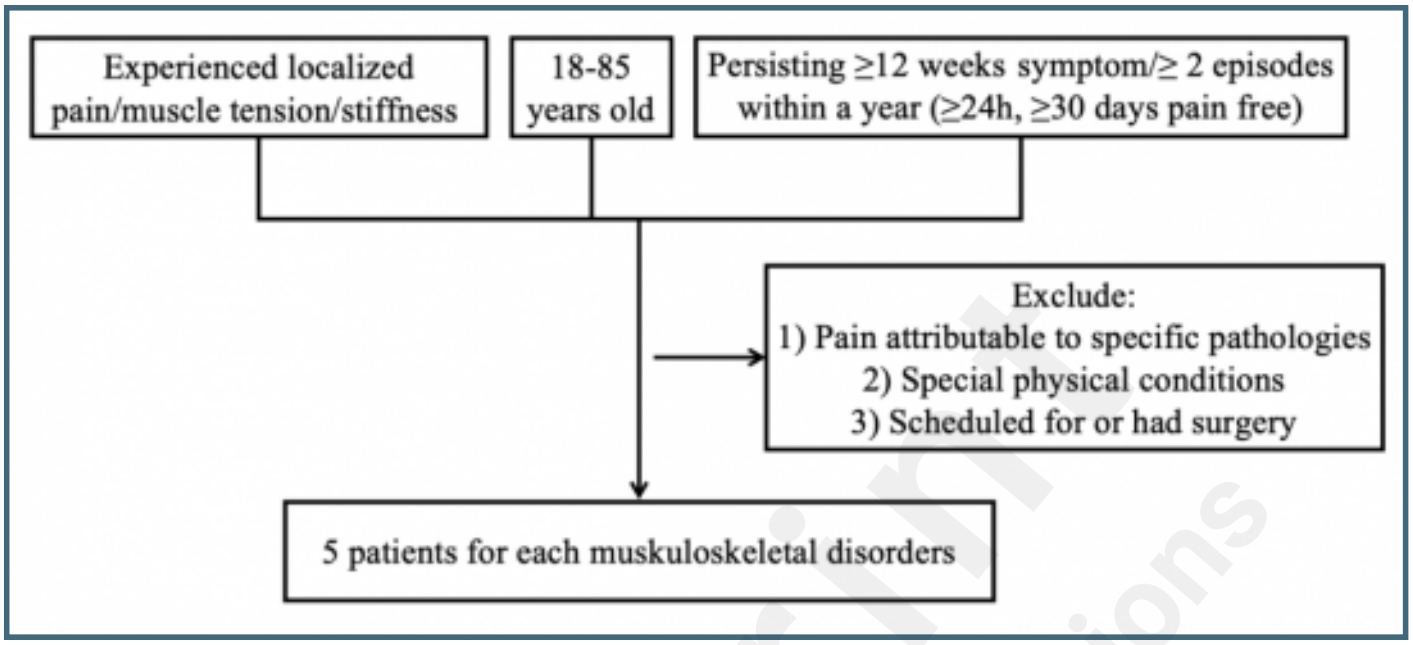
Untitled.
URL: http://asset.jmir.pub/assets/ed79a4b70c149b94446fb632b2ddd1e0.docx

Figures

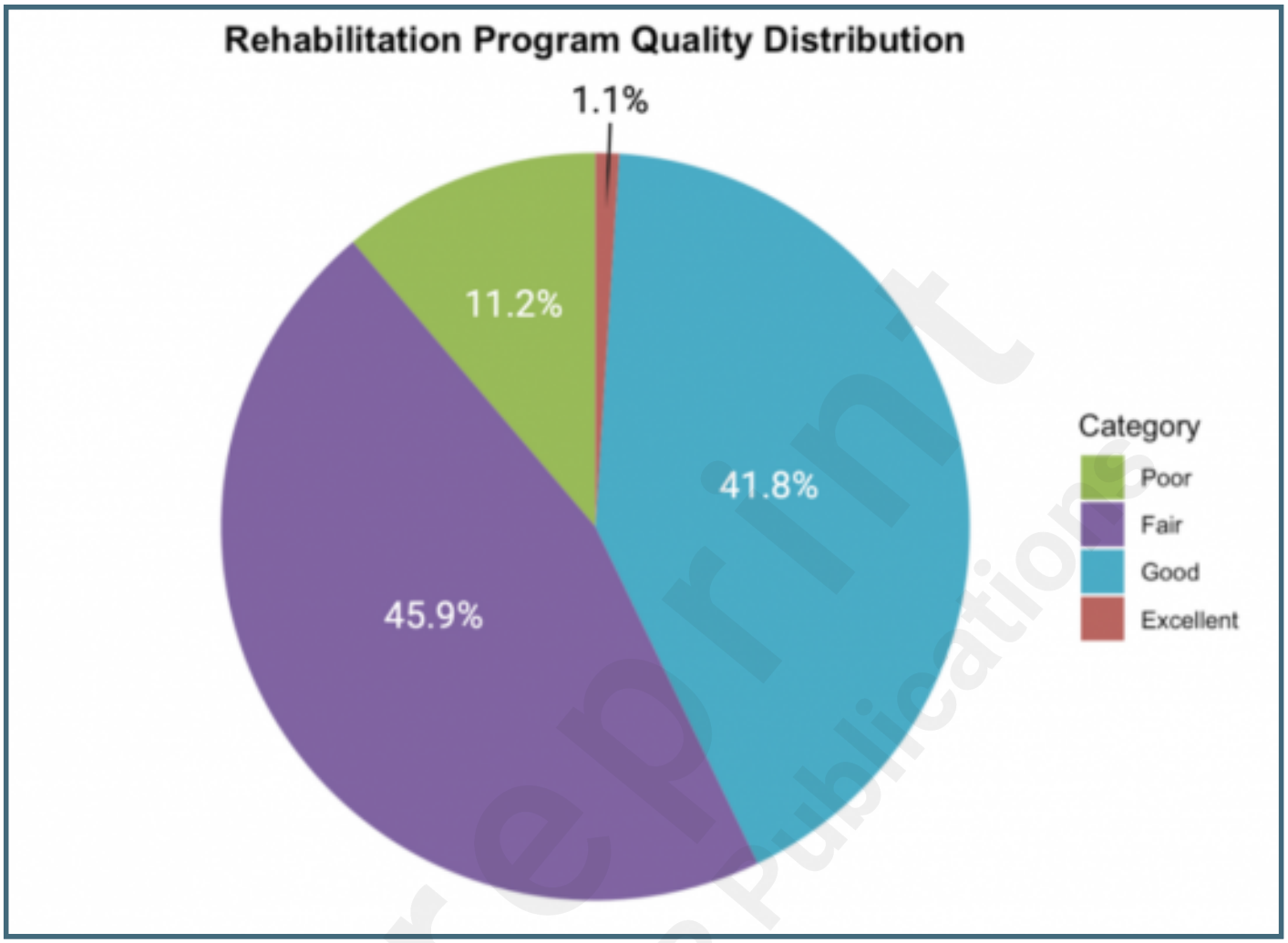
Untitled.



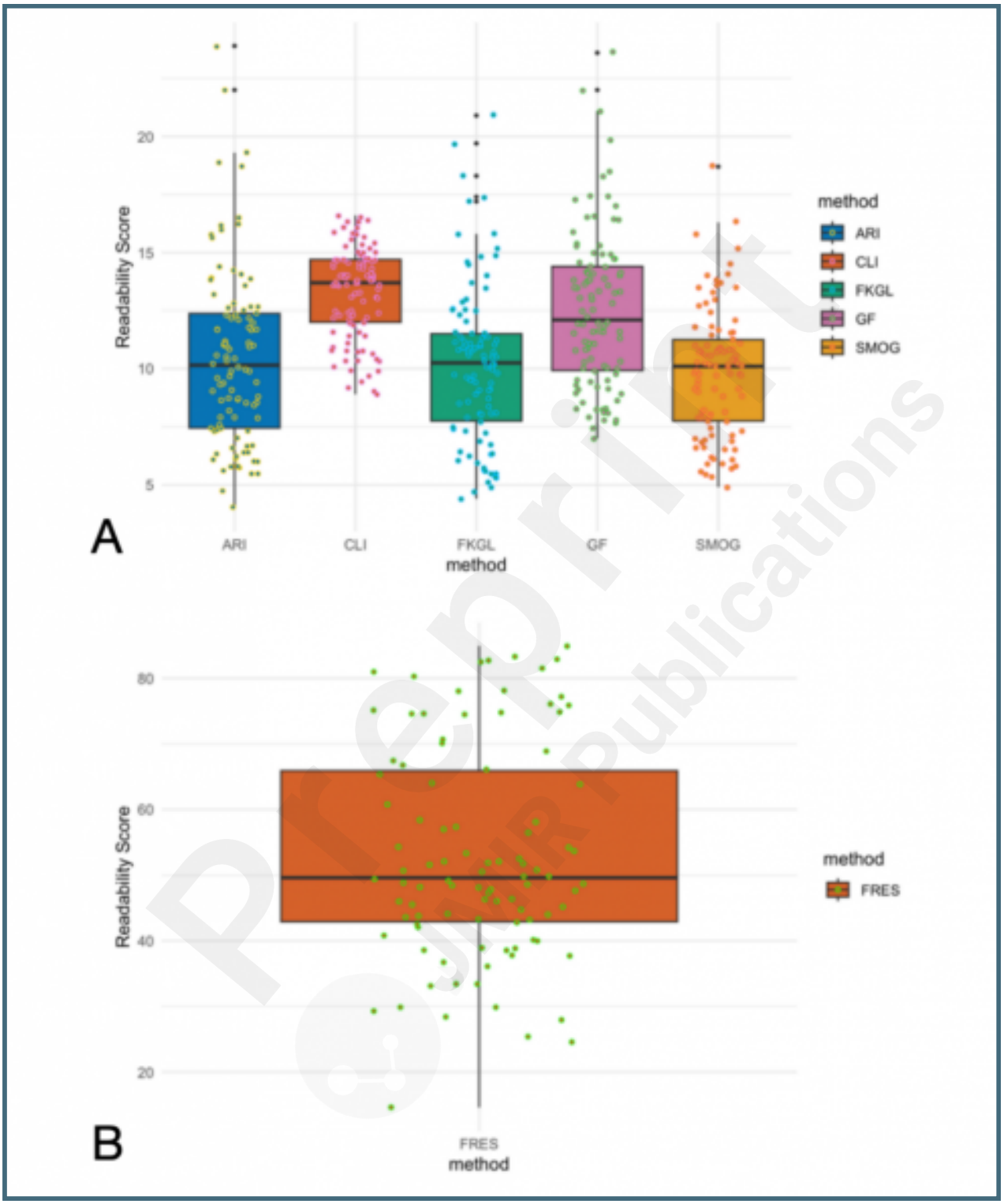
Untitled.



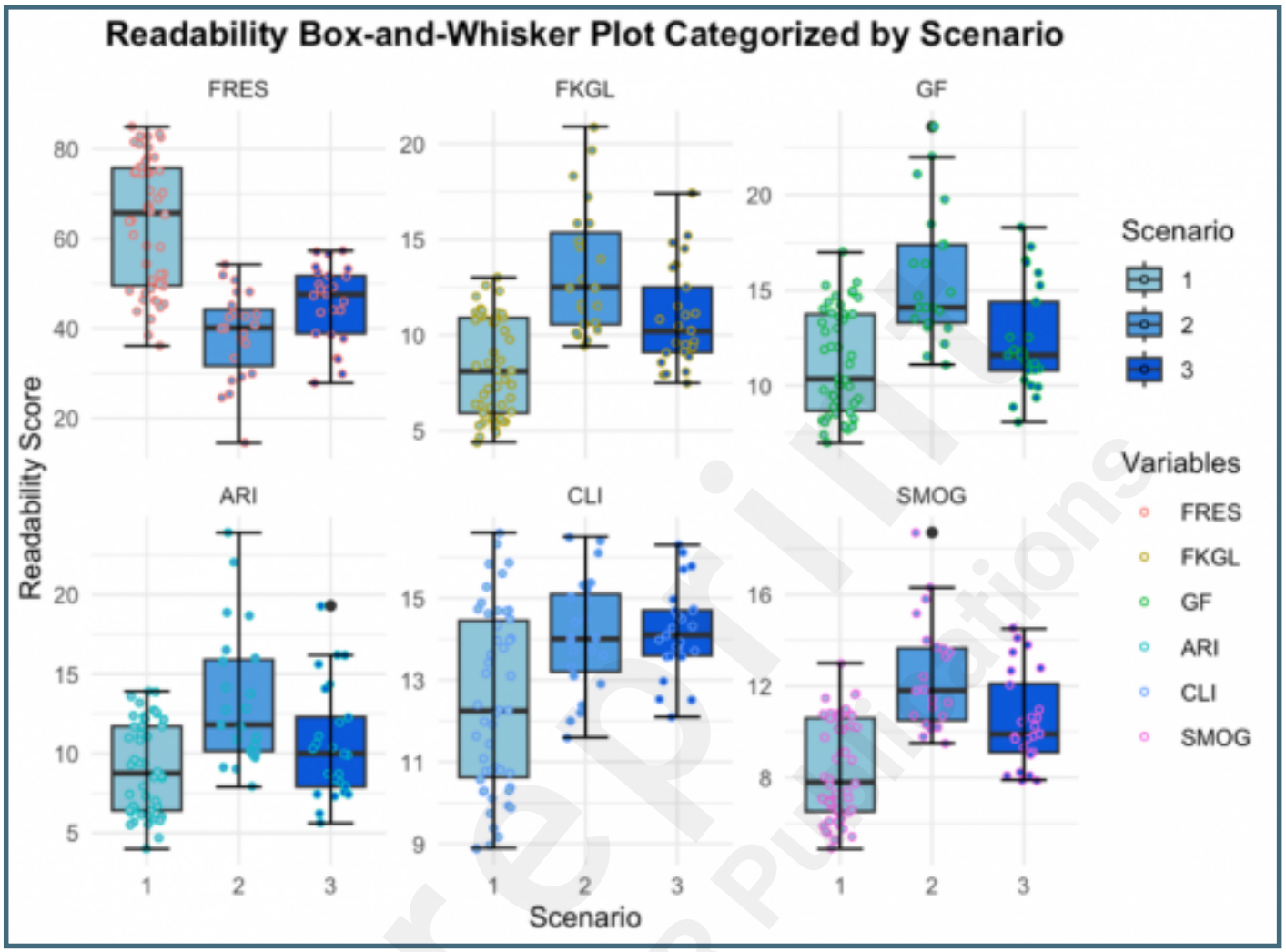
Untitled.



Untitled.



Untitled.



Untitled.