ChatGPT: a new hope for sexual dysfunction sufferers?
Metin Yigman¹, Ibrahim Untan²*, Ahmet Emin Dogan¹

¹Department of Urology, Etlik Integrated Health Campus, 06170 Ankara, Turkey
²Department of Urology, Ahi Evran University School of Medicine, 40100 Kirsehir, Turkey

*Correspondence ibrahimuntan@erciyes.edu.tr (Ibrahim Untan)

Abstract

The use of ChatGPT (Generative Pre-trained Transformer), an artificial intelligence-based language model that is on the verge of a revolutionary approach to accessing information, is rapidly increasing all over the world. The purpose of our study is evaluating the reliability and usefulness of ChatGPT responses to sexual dysfunctions, which are considered taboo in most societies. The top four Google search keywords for sexual dysfunction, erectile dysfunction, and premature ejaculation were identified for each of these disorders, and these keywords were grouped under the headings of description, causes, drug treatment, and general treatment. The content of each response was rated by two urologists for reliability and usefulness using the modified Global Quality Score rating scale, a five-point Likert scale. When the total scores of the raters were evaluated, the reliability scores were calculated as 27 for sexual dysfunction, 32 for erectile dysfunction and premature ejaculation, 29 for sexual dysfunction, 33 for erectile dysfunction, and 34 for premature ejaculation for the usefulness scores. Reliability scores showed acceptable sexual dysfunction and good sexual dysfunction agreement for erectile dysfunction and premature ejaculation (α: 0.727, 0.833, respectively). Usefulness scores showed acceptable agreement for sexual dysfunction (α = 0.727), good agreement for erectile dysfunction (α = 0.842), and questionable agreement for premature ejaculation (α = 0.600). When the scores of the raters were evaluated separately between diseases, both reliability and usefulness scores (p = 0.093 and 0.115 for reliability; p = 0.632 and 0.503 for usefulness) revealed no significant difference. ChatGPT has the potential to be a useful and reliable resource for patients to learn about sexual dysfunction, erectile dysfunction, and premature ejaculation. Artificial intelligence, whose development we cannot stop, needs improvements that can reference accepted sources and apply them to their answers to reduce potential risks and negative outcomes.

Keywords
ChatGPT; Sexual dysfunction; Erectile dysfunction; Premature ejaculation; Reliability; Usefulness

1. Introduction

John McCarthy, who contributed greatly to the development of the concept of Natural Language Processing (NLP) in the field of artificial intelligence, was among a group of scientists who defined artificial intelligence (AI) and pioneered work in this field at the Dartmouth Conference in 1956 [1]. This conference is considered the official birth of AI. Subsequent to this, AI has made substantial advances in terms of its capabilities, expanded the range of its applications, and is progressing inexorably into the future.

ChatGPT is a language model developed by OpenAI. It is trained using the GPT (Generative Pre-trained Transformer) architecture. This model has natural language understanding and generation capabilities and can be used in various language processing tasks. ChatGPT understands users’ questions and produces logical, informative answers. ChatGPT can be used in a general-purpose chat environment and is an important step forward in developing text-based systems and language understanding. ChatGPT can provide information on a variety of topics. Users can ask questions, get advice, or have general conversations on any topic. This wide range of applications has made ChatGPT attractive to users and increased its popularity. Moreover, ChatGPT can be presented with a user-friendly interface and has become generally accessible through web-based chat interfaces or applications. This has made it easy to access ChatGPT without requiring technical skills and has helped it gain popularity among users.

Online information and social media are increasingly important in healthcare, and many patients are turning to these sources for information. An analysis of data from the Health Information National Trends Survey showed that up to 80% of US adults use the Internet to search for health information for themselves [2]. Most of these sources were obtained from
non-medical peers through social media or unverified or peer-reviewed online forums [3].

The fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) defines Sexual Dysfunction (SD) (erectile dysfunction, ejaculation disorders, decreased sexual desire in men, etc.) as a significant limitation in sexual response and pleasure or pain during intercourse and states that this condition causes clinically significant distress lasting at least six months. The most common sexual dysfunctions are erectile dysfunction (ED) and premature ejaculation (PE) [4].

Sexual dysfunctions are common and their negative impact on quality of life is now widely recognized [5]. The Global Survey of Sexual Attitudes and Behaviors found that 43% of female respondents worldwide had at least one sexual problem [6]. However, very few people with sexual dysfunctions seek medical help [5]. In addition to individual factors, factors such as sociocultural, religious, economic, and the functioning of the health system may also contribute to this situation.

This study used the answers of keywords associated with SD, ED, and PE in ChatGPT to investigate their usability and reliability of them. To the best of our knowledge, no research articles about ChatGPT in the field of sexual dysfunction occur in the current literature.

2. Materials and methods

The most frequently searched keywords on Google for SD, ED, and PE related to each disease were identified utilizing Google Trends (Supplementary Fig. 1). On 09 August 2023, each disorder was logged into a separate Google trend search. The whole world, 2004 to date, and health subheadings were picked as search criteria. In the results, “top” was selected in the related questions section. In order to ascertain the most prevalent search terms associated with each disorder, the 25 most commonly sought and displayed keywords on Google were identified. These keywords were grouped under the headings of definition, causes, drug treatment, and general treatment. The topics stated for each disorder were typed in the chat zone of the ChatGPT AI chatbot. In the existing chat, the previous title and the answer given could affect the answers given to each title typed in the chat zone. Consequently, each title was written by a different user by logging into the system in question. Screenshots were taken to record each answer (Supplementary Fig. 2). The responses were generated by the 09 August 2023 version of ChatGPT. Two urologists evaluated the content of each answer. For scoring, the Global Quality Score (GQS) evaluation scale defined by Bernard et al. [7] was used, modified under two headings reliability and usefulness. The ChatGPT reliability and usefulness score is a Likert-type scale with scores ranging from 1 to 5 and it is based on the fact that the answers are from medical scientific sources, and are useful for patients (Table 1).

All statistical analyses were performed using the statistical software SPSS, version 25 (IBM, Armonk, NY, USA). Cronbach $\alpha$ and 95% confidence intervals (CI) were utilized to assess inter-rater compliance. According to intraclass correlation coefficient results, positive values $0.5 > \alpha$ indicate unacceptable agreement; $0.6 > \alpha \geq 0.5$ indicate poor agreement; $0.7 > \alpha \geq 0.6$ indicates questionable agreement; $0.8 > \alpha \geq 0.7$ indicates acceptable agreement; $0.9 > \alpha \geq 0.8$ indicate good agreement and $\alpha \geq 0.9$ indicate excellent agreement. Descriptive statistics were presented as mean $\pm$ standard deviation, median, minimum, and maximum values. Since the number of variables was less than 30, non-parametric tests were employed without assessing their normal distribution. To compare differences between groups, the Kruskal-Wallis test was utilized. $p$-values less than 0.05 were considered statistically significant.

3. Results

The results of the ChatGPT assessment of the identified topics, including reliability and usefulness scoring using a five-point Likert scale by two independent, experienced raters, are presented in Table 2. Reliability scores showed acceptable SD and good SD agreement for ED and PE ($\alpha$: 0.727, 0.833, and 0.833, respectively). Usefulness scores showed acceptable agreement for SD ($\alpha$: 0.727), good agreement for ED ($\alpha$: 0.842), and questionable agreement for PE ($\alpha$: 0.60). When the total scores of the raters were evaluated, the reliability scores were calculated as 27 for SD, 32 for ED and PE, 29 for SD, 33 for ED, and 34 for PE for usefulness scores. The mean, median, minimum, and maximum values of the scores given by the raters for the diseases are shown in Table 3. When the raters’ scores were evaluated separately between the diseases, neither reliability nor usefulness scores (0.093 and 0.115 for reliability; 0.632 and 0.503 for usefulness) showed significant differences (Table 3).

4. Discussion

The main conclusion of our research is that the ChatGPT is an acceptably useful and reliable resource for gaining knowledge about sexual dysfunction in general and its most common cause, erectile dysfunction, and premature ejaculation. Furthermore, the scales used to qualify the usefulness and reliability of the information supplied by the ChatGPT are consistent and reliable. There was a questionable degree of consistency between raters for usefulness only in PE.

Artificial intelligence technologies have vast potential to improve healthcare, optimize diagnosis and treatment processes, and improve patient care. These technologies can be employed in conjunction with medical imaging techniques to identify and diagnose diseases, assess disease risk, and forecast disease progression by analyzing vast quantities of data. This enables the development of more effective treatment plans tailored to the individual characteristics of patients. Additionally, they can assist surgeons in complex surgical procedures by integrating with surgical robots. Furthermore, they can be utilized to analyze health data and utilize it for disease management. They can also be employed to monitor disease outbreaks, construct epidemiological models, and assist in the management of health services by employing big data analytics and machine learning techniques. These are just a few examples, and artificial intelligence will find a much wider usage area with the advancing technological developments in the health field.

Epidemiologic data have shown that the prevalence and
TABLE 1. Reliability and usefulness scores.

Modified Global Quality Score (GQS) tool for reliability

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Poor reliability, most information missing</td>
</tr>
<tr>
<td>2</td>
<td>Generally poor reliability, some information listed but many important topics missing</td>
</tr>
<tr>
<td>3</td>
<td>Moderate reliability, some important information is adequately discussed but others poorly discussed</td>
</tr>
<tr>
<td>4</td>
<td>Good reliability, most of the relevant information is listed, but some topics not covered</td>
</tr>
<tr>
<td>5</td>
<td>Excellent reliability</td>
</tr>
</tbody>
</table>

Modified Global Quality Score (GQS) tool for usefulness

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Poor usefulness, not at all useful for patients</td>
</tr>
<tr>
<td>2</td>
<td>Generally poor usefulness, very limited use to patients</td>
</tr>
<tr>
<td>3</td>
<td>Moderate usefulness, somewhat useful for patients</td>
</tr>
<tr>
<td>4</td>
<td>Good usefulness, useful for patients</td>
</tr>
<tr>
<td>5</td>
<td>Excellent usefulness, very useful for patients</td>
</tr>
</tbody>
</table>

TABLE 2. Distribution, comparison, and agreement of inter-rater reliability and usefulness scores.

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Cronbach α (95% CI lower–upper)</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Cronbach α (95% CI lower–upper)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD</td>
<td>Reliability Score</td>
<td>Usefulness Score</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is SD?</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD cause</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD drug T</td>
<td>4</td>
<td>3</td>
<td>0.727 (−3.211 to 0.982)</td>
<td>3</td>
<td>4</td>
<td>0.727 (−3.211 to 0.982)</td>
</tr>
<tr>
<td>SD treatment</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total score</td>
<td>14</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ED</td>
<td>What is ED?</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>ED cause</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ED drug T</td>
<td>4</td>
<td>3</td>
<td>0.833 (−1.573 to 0.989)</td>
<td>4</td>
<td>4</td>
<td>0.842 (−1.438 to 0.99)</td>
</tr>
<tr>
<td>ED treatment</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total score</td>
<td>17</td>
<td>15</td>
<td>17</td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PE</td>
<td>What is PE?</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>PE cause</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PE drug T</td>
<td>4</td>
<td>4</td>
<td>0.833 (−1.573 to 0.989)</td>
<td>4</td>
<td>4</td>
<td>0.600 (−5.176 to 0.974)</td>
</tr>
<tr>
<td>PE treatment</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total score</td>
<td>17</td>
<td>15</td>
<td>17</td>
<td>17</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SD: sexual dysfunction; ED: erectile dysfunction; PE: premature ejaculation; T: therapy; CI: confidence intervals.
The previous month patients had searched the Internet for medical information in presenting to a general practitioner showed that 28% of patients researched their current problems before consulting. Regularly searched the Internet for medical information and support for patients with sexual dysfunction is still in its early stages, and there is limited information about the available information. In the study by Zhang et al. on help-seeking behaviors for erectile dysfunction in China, most patients consulted more than one source of information for erectile dysfunction. Doctors and the Internet were the most frequently consulted sources, and they concluded that younger patients tended to rely more on the Internet and consult more diverse sources. Previous studies have also shown the impact of social media and the unreliable content frequently shared and posted. An analysis of YouTube, a popular online video streaming service, found that medical professionals did not publish 37% of ED-related video content, and 28% contained misinformation. Similarly, Loeb et al. looked at the quality of medical information in YouTube videos on ED and found that many videos were of poor quality, 22% were trying to sell specific treatments to viewers.

As the quantity of medical data continues to grow and the complexity of clinical decision-making increases, the potential for NLP (Neuro-Linguistic Programming) tools, a branch of artificial intelligence, to assist physicians in making well-informed and timely decisions, thereby enhancing the overall quality and efficiency of healthcare, is considerable. ChatGPT fulfilled at or near the passing threshold for the United States Medical Licensing Examination without any specialized training, certifying its potential for medical education and clinical decision support.

Artificial intelligence applications can be expected to have better potential to answer more specific and clear questions. In our study, we think that the lower scores of the raters related to SD compared to ED and PE may be because SD is a top heading that includes many diseases, especially ED and PE. However, in a study evaluating ChatGPT-generated responses to 180 questions supplied by 33 physicians from 17 specialties, the median accuracy score of 180 responses was 5 (mean 4.4, SD 1.7) using a 6-point Likert scale. The result of this study demonstrated the potential of AI-based systems to procure answers to non-multiple-choice clinical questions.

The fact that there is no difference in the usefulness and reliability scores of the raters regarding the diseases indicates that they have similar knowledge about the selected topics. The most important limitation of our study is that the re-
liability of ChatGPT data is unknown, as the data source is unavailable. In addition, ChatGPT may not provide up-to-date information as it was prepared with data up to 2021. As the AI system continues to learn and develop, it may give different answers to questions written at different times. Answers were obtained by deriving general questions about diseases. Answers to more specific questions, patient-centered rather than disease-centered issues, were not evaluated.

5. Conclusions
ChatGPT is a useful and reliable resource for patients to learn about SD, ED, and PE. However, as it is a new and developing artificial intelligence application, it ought to be remembered that it may supply misinformation, which may risk patients’ health. As artificial intelligence applications continue to evolve, we think they will have more widespread and effective areas of use with more extensive and detailed studies on these issues.

AVAILABILITY OF DATA AND MATERIALS
The datasets used and analyzed during the current research are available from the corresponding author upon reasonable request.

AUTHOR CONTRIBUTIONS
MY and IU—designed the research study and wrote the manuscript; IU and AED—analyzed the data. All authors contributed to editorial changes in the manuscript. All authors read and approved the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE
Not applicable.

ACKNOWLEDGMENT
Not applicable.

FUNDING
This research received no external funding.

CONFLICT OF INTEREST
The authors declare no conflict of interest.

SUPPLEMENTARY MATERIAL
Supplementary material associated with this article can be found, in the online version, at https://onlinelibrary.wiley.com/doi/10.1002/ajmg.a.40157.

REFERENCES


