Proceedings of the 6th International Conference on Statistics: Theory and Applications (ICSTA'24)
Barcelona, Spain – August 19 - 21, 2024
Paper No. 108
DOI: 10.11159/icsta24.108

Evaluation of Biostatistics Contents in ChatGPT: A Descriptive Study

Arzu Baygül Eden[1], Alev Bakır Kayı[2], Mert Veznikli[1]
[1]Koç University
School of Medicine, Department of Biostatistics, Istanbul, Turkey
abaygul@ku.edu.tr; mveznikli@ku.edu.tr
[2]Istanbul University
Institute of Child Health, Istanbul, Turkey
alevbakirkayi@istanbul.edu.tr

Abstract **-** This study aims to evaluate the reliability and quality of ChatGPT within the context of biostatistics. The findings will enlighten researchers and clinicians about the advantages and limitations of employing ChatGPT for biostatistical information. It is important to note that this study does not extensively assess advanced biostatistical methods but rather focuses on the question: "Can researchers/clinicians dependably and effortlessly use ChatGPT?" ChatGPT was presented with Frequently Asked Questions (FAQ) in biostatistics, and responses to 20 questions were blindly evaluated by three biostatisticians holding PhDs for reliability and quality. Ratings were based on a reliability score (1 to 7), Global Quality Scale (GQS) (1 to 5), Flesch Reading Ease Score (FRES), and the Intraclass Correlation Coefficient (ICC). Moderate ICC values were observed between raters for reliability (0.646) and GQS (0.545), with a significant correlation between the reliability score and GQS ($r=0.708$; $p<0.001$). While ChatGPT provided reliable, high-quality content in response to biostatistics FAQs, it is noted that it cannot replace biostatistics experts. The readability of the content was generally challenging (FRES score: 17.2±12.04). ChatGPT shows promise as a supplementary tool for accessing biostatistics information but should be used alongside human expertise. Future research could explore ways to enhance its readability and compare its performance with alternative sources.

Keywords**:** ChatGPT, Reliability, Readability, Biostatistics.

1. Introduction

Following the Google era, Artificial Intelligence (AI) models provide a revolutionary tool to collect information. There are increasing number of studies that use Artificial Intelligence in every field of science, including medicine and biostatistics[1-7]. Now, these researchers can obtain high-quality and limitless information using AI models. One such AI model, Chat Generative Pre-trained Transformer (ChatGPT), powered by OpenAI's GPT-3.5 architecture, has gained popularity for its ability to generate human-like responses to users' queries.

ChatGPT is an AI language model developed by Open AI. ChatGPT is a free online source that generates human-like responses to chat requests using deep-learning technology[8]. ChatGPT is used as a decision-supporting and informative tool in many fields of medicine[9-11]. Researchers are also using ChatGPT to help with writing scientific content[12, 13].

Biostatistics, as a subdiscipline of statistics, provides very important techniques, applied to medicine[14].

It is obvious that using biostatistics in scientific research is important to obtain accurate and reliable information for sound research and analyses.

Researchers should always be confident in their analyzed results because they inform critical decisions and conclusions. With the increasing availability of AI-driven conversational agents like ChatGPT, there is a growing interest in exploring their potential as a resource for biostatistical information.

However, assessing the reliability and usefulness of ChatGPT is essential to ensure the integrity and validity of the information obtained. [15]While ChatGPT has been trained on a vast amount of text data, including the scientific literature, there are factors that need to be considered. These factors include the limitations of training data, the potential for biases, and the model's inability to provide real-time updates.

In the context of biostatistics, the purpose of this study is to be the first to assess ChatGPT's reliability and quality. We can learn more about the accuracy and applicability of the information offered by ChatGPT by contrasting its answers to biostatistics-related queries with recognized sources, professional judgments, and accepted statistical principles.

The results of this study will educate researchers, clinicians, and academics about the benefits and drawbacks of using ChatGPT as a source of biostatistical information. The readers must be aware that, in this study, we do not evaluate the advanced biostatistical methods in depth. In this paper, we focus on the following research question:

"Can researchers/clinicians reliably and easily use ChatGPT?"

To answer this question, the most Frequently Asked Questions (FAQ) in biostatistics were provided, and the responses were evaluated by three biostatistics experts.

In this descriptive study, we aim to say whether researchers/clinicians can use the biostatistics content from ChatGPT as an informative and supportive guide.

2. Methods

To assess the reliability and quality of ChatGPT's answers to Frequently Asked Questions about biostatistics on 27.06.2023, we used this research query: "What are the most frequently asked questions in biostatistics?" The questions were asked in English.

In response, 20 answers were retrieved (Table 1).

Table 1: Frequently asked questions in biostatistics and number of words

Questions	Words
1. What is biostatistics?	344
2. What is the role of biostatistics in healthcare and medical research?	436
3. How do I choose the appropriate study design for my research?	478
4. How do I calculate sample size for my study?	463
5. What statistical methods should I use to analyze my data?	505
6. How do I interpret p-values and confidence intervals?	459
7. What is the difference between a type I error and a type II error?	422
8. How do I assess statistical significance?	484
9. How do I perform regression analysis?	509
10. What is the difference between observational studies and experimental studies?	455
11. How do I perform survival analysis or time-to-event analysis?	465
12. What are odds ratios and how do I interpret them?	469
13. How do I conduct meta-analysis?	621
14. What are some common biases in epidemiological studies and how do I address them?	566
15. What is the difference between sensitivity and specificity?	379
16. How do I handle missing data in my analysis?	410
17. What are some best practices for data visualization in biostatistics?	513
18. How do I account for confounding variables in my analysis?	441
19. How do I analyze categorical data in biostatistics?	402
20. What are some common statistical software packages used in biostatistics and how do I use them effectively?	403

The determined questions were written in the chat box by ChatGPT. The variability in ChatGPT's responses was observed. The recorded screenshots were blindly evaluated by three investigators (ABE (Rater 1), ABK (Rater 2), and MV (Rater 3)). Three researchers rated the contents without sharing their scores with each other. All the raters are experienced biostatisticians with PhDs.

The reliability of the content of each answer was evaluated using the ChatGPT reliability score developed by Cuma Uz and Ebru Umay[16]. The scale ranged from 1 to 7 (Likert type): 1 Completely unsafe; 2: Very unsafe; 3: Relatively reliable; 4: Reliable; 5: Relatively–very reliable; 6: Very reliable; 7: Absolutely reliable.

The other aim of this study was determined as the usefulness of the contents. Global Quality Scale (GQS) [17-19] is used to measure the quality of content from different platforms, including blogs, YouTube, and websites. In this study, we

used the GQS to evaluate the quality of the contents. The GQS was scored on a 5-point scale ranging from 1 ("low quality") to 5 ("high quality"); videos were considered to be high quality (4 or 5), medium quality (3), or low quality (1 or 2).

Readability was assessed by using the Flesch Reading Ease Score (FRES) [20-23], which was calculated using a formula from an online calculator [24]. The FRES was graded from "very easy to read" to "extremely difficult to read" (Table 2).

$$FRES = 206.835 - 1.015\left(\frac{Total\ Words}{Total\ Sentences}\right) - 84.6\left(\frac{Total\ Syllables}{Total\ Words}\right) \quad (1)$$

Table 2: Interpretation of FRES

Score	Summary
90-100	Very easy to read
80-90	Easy to read
70-80	Fairly easy to read
60-70	Plain English
50-60	Fairly difficult to read
30-50	Difficult to read
10-30	Very difficult to read
0-10	Extremely difficult to read

The study adheres to the national rules in the field and ethical principles outlined in the Helsinki Declaration.

2.1. Statistical Analysis

All statistical analyses were performed by using IBM Corp. (released 2021; IBM SPSS Statistics for Windows, Version 28.0. Armonk, NY: IBM Corp.), and $p<0.05$ was considered to be significant.

Descriptive statistics were presented by using the mean, standard deviation, median, minimum value, and maximum value, and 25th and 75th quarters. Normality assumption was checked by using Shapiro Wilks test. The Intraclass Correlation Coefficient (ICC) was used to determine concordance between the raters. The Friedman Test was performed to find the change between the scores obtained by the raters. For pairwise post hoc evaluation, Wilcoxon Signed Rank test with Bonferroni Correction was used. Spearman's Rho Correlation coefficient was applied to examine the pairwise relationship between the raters.

2.2. Ethical Consideration

It was not necessary to obtain ethical approval because the study did not use any human participants.

3. Results

A total of 20 questions were evaluated by three raters using the reliability score (from 1 to 7) and Global Quality Scale (GQS) (from 1 to 5). Descriptive statistics of three raters' answers are shown in Table 3.

Table 3: Descriptive statistics for raters' answers and FRES

	Rater 1 (A.B.E.)		Rater 2 (A.B.K.)		Rater 3 (M.V.)		
	Reliability Score	**GQS**	**Reliability Score**	**GQS**	**Reliability Score**	**GQS**	**FRES**
Mean	6.2	4.6	5.8	4	5.4	4.3	17.21
Standard Deviation	1.005	0.754	0.639	0.562	1.0	0.6	12.04
Median	6.5	5	6	4	6	4	18.02
Minimum-Maximum	4-7	3-5	4-7	3-5	3-7	3-5	0.54-44.73
Percentiles (25-75)	6-7	4.25-5	5.25-6	4-4	5-6	4-5	8.13-23.27

GQS: Global Quality Scale, FRES: Flesch Reading Ease Score

The minimum value for the mean of three raters (for 20 questions) was 4.3, whereas the maximum value was 6.7 for the reliability score, (minimum was 3, and maximum was 4.7 for the GQS).

Readability assessed using the Flesch Reading Ease Score for 20 questions, which was found to be 17.21+12.04, and the range was from 0.54 to 44.73.

3.1. Agreement and Comparison of Raters

The Intraclass Correlation Coefficient (ICC) was calculated between three raters. A moderate ICC was found for the reliability and GQS scores (0.646-0.599) (Table 4).

Table 3: Agreement statistics between raters' answers

	Reliability Score	**GQS**
Intraclass Correlation Coefficient	0.646	0.599
Spearman's rho Correlation Test, p, r		
Rater1-Rater2	0.002, 0.638	0.021, 0.511
Rater1-Rater3	0.083, 0.397	0.219, 0.287
Rater2-Rater3	0.032, 0.482	0.578, 0.132

GQS: Global Quality Scale

There was statistically significant difference between the raters in terms of the reliability score (Friedman test; p=0.002). According to the post hoc evaluation (Wilcoxon Signed Rank test with Bonferroni Correction), the difference in the reliability scores came from A.B.E (Rater 1) and M.V. (Rater 3) ($p_{ABE\text{-}ABK}$=1.00/$p_{ABE\text{-}MV}$=0.013/$p_{ABK\text{-}MV}$=0.173).

There was also a statistically significant difference between the raters in the GQS. The post hoc pairwise evaluation (Wilcoxon Signed Rank test with Bonferroni Correction) results indicated that there is difference in the reliability scores from A.B.E (Rater 1) and M.V. (Rater 3) ($p_{ABE\text{-}ABK}$=0.013/$p_{ABE\text{-}MV}$=0.291/$p_{ABK\text{-}MV}$=0.291).

The correlation between the raters was evaluated. There is a moderate and statistically significant correlation between Rater 1 and Rater 2, a weak correlation between Rater 2 and Rater 3 in terms of the reliability score, while there is no statistically significant correlation between Rater 1 and Rater 3.

The correlation between Rater 1 and Rater 2 was statistically significant and was moderate for the GQS; however, the Rater 1–Rater 3 and Rater 2–Rater 3 correlations were not statistically significant.

A positive, high-level, and statistically significant correlation ($r=0.708$; $p<0.001$) was calculated between the average reliability score and Global Quality Scales.

3.2. Readability

The answers, which were produced by ChatGPT, were evaluated by using the FRES score. The mean of the FRES score was 17.2±12.04, which shows us that all the contents were generally very difficult to read (a college graduate would be needed).

4. Discussion

The most important result of our study is that ChatGPT provides reliable and high-quality information in response to the main and most Frequently Asked Questions on biostatistics by researchers and clinicians, whereas the contents were generally "very difficult to read". In other words, researchers should be, at least, at a graduate level to better understand the contents.

The highest mean score in terms of reliability was 6.7. Thirteen of the twenty questions received a score higher or equal to six, whereas seven of them received an average score lower than six. This shows that the reliability of the contents was quite high.

Minimum reliability scores were obtained for the question numbered 5 ("What statistical methods should I use to analyze my data?") and question no 19 ("How do I analyze categorical data in biostatistics?").

The reason for obtaining the lowest score for the "What statistical methods should I use to analyze my data?" content may be due to the nature of the question. The answer from ChatGPT can be thought of as a guide; however, there are, of course, many evolving methods (and the number is still increasing) in statistics. In addition, the content did not refer to "sample size calculations", which denoted a lack of information.

The answer to the "How do I analyze categorical data in biostatistics?" question was adequate; however, some explanatory information should be given to reflect more reliability.

Questions 2, 10, 12, 13, 15, 16, and 17 received the highest (4.7) GQS. The lowest GQS was obtained for the 18th question ("How do I analyze categorical data in biostatistics?"). The mean GQS was 4.3±0.5. Nineteen out of twenty (95%) questions gained more than four points, which displays the high quality of the contents.

ChatGPT reflected minimum reliability and quality scores for the question, "How do I analyze categorical data in biostatistics?"

4.1. Limitations

There are several limitations to this study.

- First, if we query ChatGPT at different times, we do not always obtain the same results. Therefore, the contents can vary from time to time.
- ChatGPT may not always produce up-to-date information. We retrieved an answer from ChatGPT to the following query in order to obtain information about the time interval for sourcing data: "what is the cut off time for the training of Chat GPT data". The answer was "For the GPT-3.5 architecture, my training data goes up until September 2021" (retrieved July 15, 2023, from a ChatGPT conversation).
- We only retrieved answers from ChatGPT-3.5. GPT-4 may produce more detailed information.

4.2. Comparison with Prior Work

There are some studies that evaluate ChatGPT-produced content in the medical field [25-28]; however, our study is the first to evaluate the reliability and readability of biostatistics content.

At one time, Google and YouTube were thought to be the most important internet sources for science, in particular, biostatistics. Nowadays, the trend is changing, and Artificial Intelligence (AI) tools, especially ChatGPT, are becoming very popular. There is one study in the literature, in which the authors studied the reliability and usefulness of ChatGPT in Rheumatology [16]. However, there is no study about biostatistical topics. In the previous literature, there is a study about the reliability of the use of YouTube [29]. Baygul et.al. stated that 75.4 % of the included studies were classified as "useful". The mean of Global Quality Scale was 3.9±1.1. In this study, the researchers stated that the YouTube videos may be referred

to students, and most of the videos were classified as useful. Similar to this research, our study stated that ChatGPT's biostatistics content is reliable and of an acceptable quality.

Similar to Palal et. al. study, our results showed that the contents in biostatistics can be used as a guide, however researchers should use ChatGPT responsibly and rationally by approving the information from other sources[30].

Authors should discuss the results and how they can be interpreted from the perspective of previous studies and of the working hypotheses. The findings and their implications should be discussed in the broadest context possible. Future research directions may also be highlighted.

It is evident that the advancement of AI-driven virtual assistants like ChatGPT, aimed at aiding patients in managing their health, has the potential to constitute a significant advancement in the realm of medicine. These virtual assistants have the capacity to generate automated summaries of patient interactions and medical backgrounds, streamlining the task of medical record management for healthcare practitioners. Through voice dictation, healthcare professionals can harness ChatGPT's ability to automatically condense critical information, including symptoms, diagnoses, treatments, and glean pertinent data from patient records such as laboratory findings or imaging reports. Despite these advantages, the use of ChatGPT and other artificial intelligence (AI) tools in medical writing raises ethical, legal and reliability concerns. Besides copyright infringement, there is the potential for inaccuracies in the content generated by the virtual assistant or the potential for bias through misinterpretation of data. Therefore, it is important to carefully consider the limitations and issues related to the use of AI applications in medicine[10].

5. Conclusion

In conclusion, the answers to the Frequently Asked Questions in biostatistics provided by ChatGPT were reliable and were of good quality. Although ChatGPT may be used as a guide and decision-support mechanism in situations where biostatistical information is needed, it has been seen that it cannot replace biostatistics experts.

For advanced information or more technical biostatistical information, other sources should also be used. Although the readability scores of the contents are low, considering that the researchers are at least at an undergraduate level, it can be said that this is acceptable, but needs improvement. Future studies may compare the reliability, usefulness, and quality of ChatGPT alongside other sources like YouTube/Google.

Acknowledgements

Chat GPT was used in the design and conducting of this study.

Abbreviations

AI: Artificial Intelligence;
ChatGPT: Chat generative pre-trained transformer;
FAQ: Frequently asked questions;
FRES: Flesch reading ease score;
GQS: Global quality scale;
ICC: Intraclass correlation coefficient.

References

[1] R. Gupta, D. Srivastava, M. Sahu, S. Tiwari, R. K. Ambasta, and P. Kumar, "Artificial intelligence to deep learning: machine intelligence approach for drug discovery," *Molecular diversity*, vol. 25, pp. 1315-1360, 2021, doi: https://doi.org/10.1007/s11030-021-10217-3.

[2] A. N. Ramesh, C. Kambhampati, J. R. Monson, and P. J. Drew, "Artificial intelligence in medicine," (in eng), *Ann R Coll Surg Engl*, vol. 86, no. 5, pp. 334-8, Sep 2004, doi: Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1964229/.

[3] P. Hamet and J. Tremblay, "Artificial intelligence in medicine," *Metabolism*, vol. 69, pp. S36-S40, 2017, doi: https://doi.org/10.1016/j.metabol.2017.01.011.

[4] Y. Mintz and R. Brodie, "Introduction to artificial intelligence in medicine," *Minimally Invasive Therapy & Allied Technologies*, vol. 28, no. 2, pp. 73-81, 2019, doi: https://doi.org/10.4103%2Fjfmpc.jfmpc_440_19.

[5] D. A. Hashimoto, E. Witkowski, L. Gao, O. Meireles, and G. Rosman, "Artificial intelligence in anesthesiology: current techniques, clinical applications, and limitations," *Anesthesiology*, vol. 132, no. 2, pp. 379-394, 2020, doi: https://doi.org/10.1097/ALN.0000000000002960.

[6] I. Tunali, R. J. Gillies, and M. B. Schabath, "Application of radiomics and AI for lung cancer precision medicine," *Cold Spring Harbor perspectives in medicine*, vol. 11, no. 8, 2021, doi: https://doi.org/10.1101/cshperspect.a039537.

[7] B. Acs, M. Rantalainen, and J. Hartman, "Artificial intelligence as the next step towards precision pathology," *Journal of internal medicine*, vol. 288, no. 1, pp. 62-81, 2020, doi: https://doi.org/10.1111/joim.13030.

[8] E. Porter, M. Murphy, and C. O'Connor, "Chat GPT in dermatology: progressive or problematic?," *Journal of the European Academy of Dermatology and Venereology*, 2023, doi: https://doi.org/10.1111/jdv.19174.

[9] M. Sallam, "ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns," *Healthcare*, vol. 11, no. 6, p. 887, 2023, doi: https://doi.org/10.3390/healthcare11060887.

[10] T. Dave, S. A. Athaluri, and S. Singh, "ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations," *Frontiers in Artificial Intelligence*, vol. 6, p. 1169595, 2023, doi: https://doi.org/10.3389/frai.2023.1169595.

[11] A. M. DiGiorgio and J. M. Ehrenfeld, "Artificial intelligence in medicine & ChatGPT: de-tether the physician," *Journal of Medical Systems*, vol. 47, no. 1, p. 32, 2023, doi: https://doi.org/10.1007/s10916-023-01926-3.

[12] M. Salvagno, F. S. Taccone, and A. G. Gerli, "Can artificial intelligence help for scientific writing?," *Critical care*, vol. 27, no. 1, pp. 1-5, 2023, doi: https://doi.org/10.1186/s13054-023-04380-2.

[13] H. Alkaissi and S. I. McFarlane, "Artificial hallucinations in ChatGPT: implications in scientific writing," *Cureus*, vol. 15, no. 2, 2023, doi: https://doi.org/10.7759/cureus.35179.

[14] K. J. Lee, M. Moreno-Betancur, J. Kasza, I. C. Marschner, A. G. Barnett, and J. B. Carlin, "Biostatistics: a fundamental discipline at the core of modern health data science," *The Medical Journal of Australia*, vol. 211, no. 10, p. 444, 2019, doi: https://doi.org/10.5694/mja2.50372.

[15] S. H. Park, K.-H. Do, S. Kim, J. H. Park, and Y.-S. Lim, "What should medical students know about artificial intelligence in medicine?," *Journal of educational evaluation for health professions*, vol. 16, 2019.

[16] C. Uz and E. Umay, "'Dr ChatGPT': Is it a reliable and useful source for common rheumatic diseases?," *International Journal of Rheumatic Diseases*, 2023, doi: https://doi.org/10.1111/1756-185X.14749.

[17] C. M. Etzel, S. L. Bokshan, T. A. Forster, and B. D. Owens, "A quality assessment of YouTube content on shoulder instability," *The Physician and Sportsmedicine*, vol. 50, no. 4, pp. 289-294, 2022, doi: https://doi.org/10.1080/00913847.2021.1942286.

[18] C. Arslan, E. C. Aksahin, R. B. Nur Yılmaz, and D. Germec Cakan, "Does YouTubeTM Offer High-Quality Information About Nasoalveolar Molding?," *The Cleft Palate Craniofacial Journal*, vol. 61, no. 1, pp. 5-11, 2024, doi: https://doi.org/10.1177/10556656221115025.

[19] O. Yapici and E. Akman, "Evaluation of the quality, reliability and content of YouTube™ videos related to the Crimean-Congo hemorrhagic fever," *Eur Rev Med Pharmacol Sci*, vol. 26, no. 20, pp. 7719-7723, 2022, doi: https://doi.org/10.26355/eurrev_202210_30049.

[20] M. E. Chowdhury *et al.*, "Can AI help in screening viral and COVID-19 pneumonia?," *Ieee Access*, vol. 8, pp. 132665-132676, 2020, doi: https://doi.org/10.1371/journal.pone.0257884.

[21] H. Raja and N. Fitzpatrick, "Assessing the readability and quality of online information on Bell's palsy," *The Journal of Laryngology & Otology*, pp. 1-5, 2022, doi: https://doi.org/10.1017/S0022215122002626.

[22] D. Ghanem, O. Covarrubias, A. B. Harris, and B. Shafiq, "Readability of the Orthopaedic Trauma Association Patient Education Tool," *Journal of Orthopaedic Trauma*, 2023, doi: https://doi.org/10.1097/BOT.0000000000002593.

[23] C. A. Martin, S. Khan, R. Lee, A. T. Do, J. Sridhar, E. L. Crowell, and E. C. Bowden, "Readability and suitability of online patient education materials for glaucoma," *Ophthalmology Glaucoma*, vol. 5, no. 5, pp. 525-530, 2022, doi: https://doi.org/10.1097/IJG.0000000000002012.

[24] "FRES Calculator 2023." https://goodcalculators.com/flesch-kincaid-calculator/ (accessed Aug. 2, 2023).

[25] S. Liu *et al.*, "Using AI-generated suggestions from ChatGPT to optimize clinical decision support," *Journal of the American Medical Informatics Association*, vol. 30, no. 7, pp. 1237-1245, 2023, doi: https://doi.org/10.1093/jamia/ocad072.

[26] A. M. Fayed, N. S. B. Mansur, K. A. de Carvalho, A. Behrens, P. D'Hooghe, and C. de Cesar Netto, "Artificial intelligence and ChatGPT in Orthopaedics and sports medicine," *Journal of Experimental Orthopaedics*, vol. 10, no. 1, pp. 1-7, 2023, doi: https://doi.org/10.1186/s40634-023-00642-8.

[27] G. Currie and K. Barry, "ChatGPT in nuclear medicine education," *Journal of Nuclear Medicine Technology*, 2023, doi: https://doi.org/10.2967/jnmt.123.265844.

[28] R. Golan *et al.*, "ChatGPT's ability to assess quality and readability of online medical information: evidence from a cross-sectional study," *Cureus*, vol. 15, no. 7, 2023.

[29] A. B. EDEN and N. G. İNAN, "Are YouTubeTM Videos Useful for Biostatistics Education: A sample of Logistic Regression," *Clinical and Experimental Health Sciences*, vol. 12, no. 4, pp. 840-844, 2022, doi: https://doi.org/10.33808/clinexphealthsci.1058931.

[30] D. Palal, S. Ghonge, V. Jadav, and H. Rathod, "ChatGPT: a double-edged sword?," *Health Services Insights*, vol. 16, p. 11786329231174338, 2023.