Journal of Community Medicine & Health Education

ISSN: 2161-0711

Research Article Open Access

A Quality Analysis of Laparoscopic Donor Nephrectomy-related Information Disseminated by Artificial Intelligence Chatbots using Validated Tools

Matthew Wainstein[1]*, Isaac DeMoss[1], Stephen Hong[1], Mehdi Nayebpour[2], Naoru Koizumi[2], Obi Ekwenna[1]

[1]*Department of Urology, The University of Toledo Heath Science Campus, USA*
[2]*Schar School of Policy and Government, George Mason University, USA*

Abstract

Background: Artificial intelligence (AI) chatbots, such as ChatGPT and Bard, have become popular sources of medical information and are likely to be used by potential kidney donors seeking information. Despite their potential role in guiding patients' inquiries, the ability of AI chatbots to provide quality information still needs to be further investigated. This study aims to assess and compare the quality of donor nephrectomy-related information provided by ChatGPT and Bard.

Methods: A set of questions regarding kidney donation was generated based on general information from the National Kidney Foundation and the United Network for Organ Sharing. The questions were then typed directly into ChatGPT and Google Bard, and the responses were recorded and assessed for eligibility criteria. Three reviewers utilized two validated tools for evaluating health information, the DISCERN and PEMAT-P tools, to grade information quality, understandability, and actionability.

Findings: A total of 40 of 42 screened responses were included in the study, with two responses excluded for not containing information relevant to donor nephrectomies. There were no significant differences between ChatGPT and Bard based on assessment with the DISCERN, PEMAT-P Understandability, and PEMAT-P Actionability tools. Performance on the DISCERN and PEMAT-P Actionability surveys was notably poor, while most of the responses were "understandable" based on the PEMAT-P Understandability tool.

Interpretation: Both ChatGPT and Bard provide relevant and understandable responses. However, the quality of information is generally poor, and neither chatbot provides "actionable" responses. While AI chatbots have the potential for use in responding to donor nephrectomy-related queries, caution should be used.

Keywords: Donor nephrectomy; Laparoscopic; Artificial intelligence; Chatbot; ChatGPT; BARD

Introduction

Live donor nephrectomies are unique surgical procedures as they are performed on completely healthy patients solely for the benefit of another person. Live donor nephrectomies are considered safe, have few complications, and generally lead to good outcomes [1-3]. Since the first live donor nephrectomy was performed in 1954, the surgical practice has evolved significantly, with a focus on increasing donor safety and quality of life [1-4]. As the surgical practice has shifted away from open surgery and towards minimally-invasive techniques, laparoscopic surgery has become the predominant method for transplant surgeons [1-5]. This has led to fewer complications, shorter hospital stays, and quicker donor recovery [6,7]. In 2023, there was 21,764 total kidney donations, 6,293 of which have been from live donors [8]. In contrast, there are 96,012 patients currently on the kidney donation waiting list [8]. While there has been an increase in both deceased and cumulative kidney donations over the last decade, the number of live donors has been variable each year [8]. Furthermore, the waitlist is continuing to grow faster than donor recruitment and transplantation rates [9]. The extreme shortage of kidneys available for transplant has become a crisis for patients with end-stage renal disease, and increasing the number of live kidney donors is one of few ways to address this problem [10]. Donating a kidney requires deliberate thought and knowledge about the eligibility process and surgical procedure. The internet is a widely accessible platform for patients to supplement their medical knowledge and decision-making [11]. Specifically, Artificial Intelligence (AI) chatbots, such as ChatGPT and Bard, have gained widespread internet popularity. They are openly accessible, easy to use, and can provide concise answers to specific questions in a matter of seconds. For these reasons, it is reasonable to assume that potential kidney donors may turn to AI chatbots to receive information. This calls into question the quality of information provided by AI chatbots. There are many common myths and misconceptions regarding live kidney donation, and misinformation could deter prospective kidney donors [12]. Despite their potential influence over patients' health queries, the accuracy and efficacy of medical information provided by AI chatbots is understudied. This study aims to analyze and compare the quality of the information provided by ChatGPT and Google Bard regarding laparoscopic donor nephrectomies

*Corresponding author: Matthew Wainstein, Department of Urology, The University of Toledo Medical Center, USA, Email: matthew.wainstein@rockets.utoledo.edu

Received: 14-June-2024, Manuscript No. jcmhe-24-138954; **Editor assigned:** 17-June-2024, PreQC No. jcmhe-24-138954 (PQ); **Reviewed:** 01-July-2024, QC No. jcmhe-24-138954; Revised: 08-July-2024, Manuscript No. jcmhe-24-138954 (R); **Published:** 15-July-2024, **DOI**: 10.4172/2161-0711.1000878

Citation: Wainstein M, DeMoss I, Hong S, Nayebpour M, Koizumi N, et al. (2024) A Quality Analysis of Laparoscopic Donor Nephrectomy-related Information Disseminated by Artificial Intelligence Chatbots using Validated Tools. J Community Med Health Educ. 13: 878.

and related questions about live kidney donation.

Methods

Search strategy and eligibility screening

A set of 21 common questions asked by potential kidney donors was developed based on general information from the National Kidney Foundation and the United Network for Organ Sharing. The questions were formatted to mimic the language and vocabulary of an actual patient and were placed into the following categories based on the nature of the question: Prospective-Questions about eligibility and preparation (e.g., How long does it take to be evaluated to be a living kidney donor?) Technical-Questions detailing the procedure (e.g., I am undergoing a laparoscopic donor nephrectomy. What does the surgery entail?) Recovery, complications, risks-Questions focusing on post-procedure aspects (e.g., How long does it take to recover after a laparoscopic donor nephrectomy?) Other-Miscellaneous questions that don't fit into the above categories. (e.g., Where can I get more information about live kidney donation?) The questions were then typed directly into ChatGPT (OpenAI, San Francisco, CA) and Bard (Google, Mountain View, CA) [13,14]. Because ChatGPT and Bard tailor responses based on prior conversations, a new conversation was generated for each question. One author (MW) recorded and assessed the responses for eligibility. Responses were excluded if they were duplicates, irrelevant to donor nephrectomies within reason, and describing partial or total nephrectomies for purposes other than transplant. The remaining responses were considered for evaluation and scoring.

Response review

Two medical students (MW and ID) and one urological resident (SH) examined the included responses using two validated tools for evaluating health information: The DISCERN tool for assessing quality of health information and the Patient Education Materials Assessment Tool for Printed materials (PEMAT-P) for assessing information quality, understandability, and actionability. The DISCERN tool is a standardized survey that untrained laypersons and health professionals can use to assess the quality of consumer health information [15,16]. The 16-question survey consists of 8 questions regarding the reliability and sourcing of the information, 7 questions focusing on specific details of information regarding treatment choices, and 1 question about the overall quality rating. Each question is scored between 1 and 5, with 5 defined as a definite "yes" in accomplishing the goal of the question, 1 as a definite "no," and 2-4 indicating partial accomplishment. The score from each category, except for the rater's overall evaluation, is summed), giving a score between 15 and 75. Scores between 63 and 75 are considered as "excellent," 51-62 as "good," 39-50 as "fair," 28-38 as "poor," and <27 as "very poor." PEMAT-P is another standardized survey that can be used to assess the understandability and actionability of printed health information [17,18]. The PEMAT-P consists of 24 questions, with the first 17 assessing understandability and the last 7 assessing actionability. Each question is scored as either 0 or 1, with 0 being "no" and 1 being "yes." Responses are considered "understandable" if at least 70% of the understandability items are met and "actionable" if at least 70% of the actionability items are met. 7 total items from the PEMAT-P survey were excluded as they were not applicable. Five items were excluded because neither chatbot provides any form of visual aid. One item was excluded because no response from either chatbot required calculations. One item was excluded because the average length of responses provided by both chatbots is "very short" (two or fewer paragraphs and no more than 1 page in length) [17].

Interrater reliability

To measure the degree of agreement between reviewers, interrater reliability (IRR) was calculated in Microsoft Excel using the percent of absolute agreement. IRR was calculated for DISCERN, PEMAT-P Understandability, and PEMAT-P actionability and is shown in Table 1. IRR scores for the PEMAT-P Understandability and Actionability tools indicated "substantial agreement" among all reviewers [19]. The IRR scores for the DISCERN tool were lower in comparison, indicating "fair agreement" [19]. We believe this is due to the partial scoring allowed by the DISCERN survey.

Table 1: Interrater Reliability for the DISCERN, PEMAT-P understanbility, and PEMAT-P actionability between each evaluator

Rater	Discern	PEMAT-P understandability	PEMAT-P actionability
MW-ID	0.63	0.89	0.87
MW-SH	0.50	0.85	0.80
ID-SH	0.53	0.84	0.85

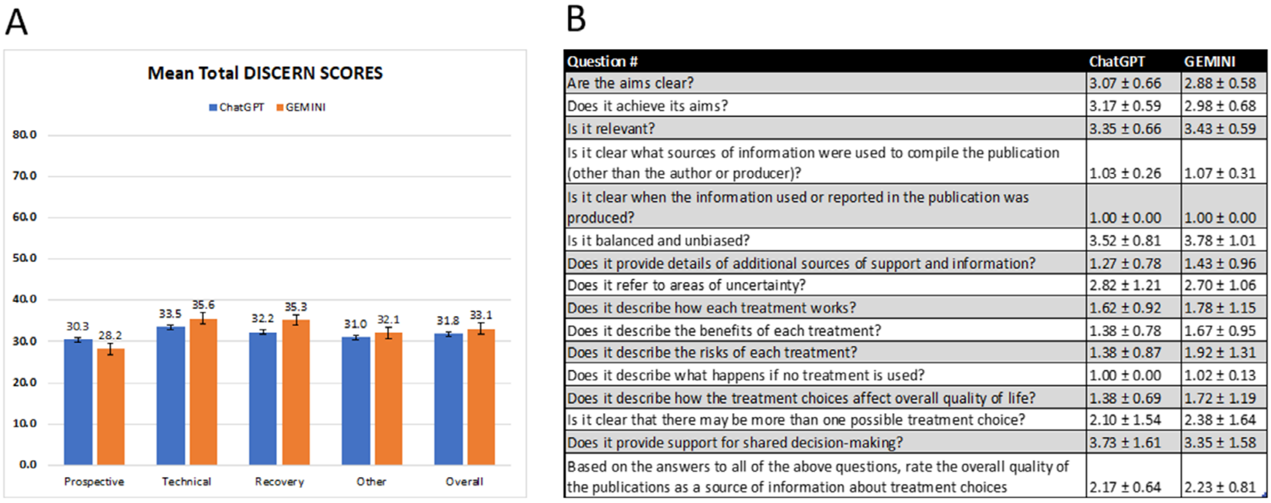
Results

General characteristics

40 of the 42 generated responses were included in the study, while two were excluded because they were irrelevant to donor nephrectomies. The average length of each response was 266 words and 277 words for ChatGPT and Bard, respectively. 31 of the 40 (77•5%) responses recommended further discussion or consultation with a physician or transplantation team. However, only three responses disclosed that it cannot provide medical advice. 17•5% (7/40) responses utilized persuasive language to describe the act of kidney donation, including phrases such as "selfless act," "gift of life," and "best thing they have ever done." Of note, 83•3% (5/6) of these responses were provided by Bard (Table 1).

Discern

The mean total DISCERN scores with standard deviations for each category are represented in Figure 1. There were no significant differences between total DISCERN scores by ChatGPT or Bard, and the overall performance was low. The scores were highest on responses to questions in the technical category. However, the mean total scores for each question category overall received a "poor" rating. Individually, 92•5% (37/40) of the responses received a "poor" rating, and 5•0% (2/40) received a "very poor" rating. Only one response from Bard received a "fair" rating, and no response from either chatbot received an "excellent" or "good" rating. The individual DISCERN questions and statistics are shown in Figure 1. The overall performance was also low. The highest performing question was 6, “Is it balanced and unbiased?” The scores were similarly high for questions pertaining to aims, relevancy, and support for shared decision-making. The lowest performing questions were 4, “Is it clear what sources of information were used to compile the publication (other than the author or producer)?” and 5, “Is it clear when the information used or reported in the publication was produced?” The scores were also very low for questions on risks and outcomes if no treatment is used.

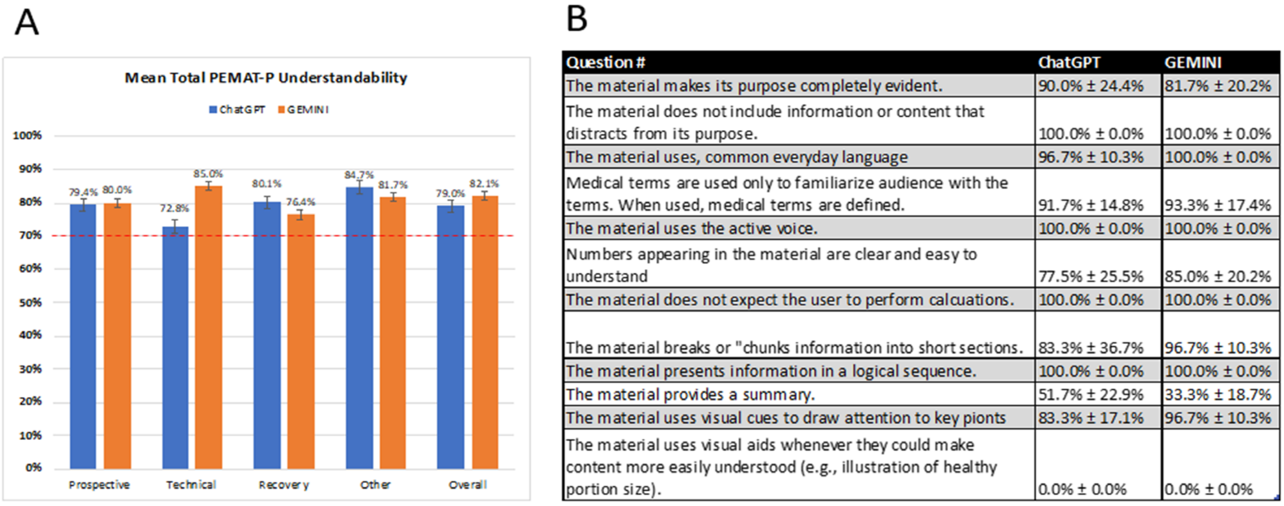


Question #	ChatGPT	GEMINI
Are the aims clear?	3.07 ± 0.66	2.88 ± 0.58
Does it achieve its aims?	3.17 ± 0.59	2.98 ± 0.68
Is it relevant?	3.35 ± 0.66	3.43 ± 0.59
Is it clear what sources of information were used to compile the publication (other than the author or producer)?	1.03 ± 0.26	1.07 ± 0.31
Is it clear when the information used or reported in the publication was produced?	1.00 ± 0.00	1.00 ± 0.00
Is it balanced and unbiased?	3.52 ± 0.81	3.78 ± 1.01
Does it provide details of additional sources of support and information?	1.27 ± 0.78	1.43 ± 0.96
Does it refer to areas of uncertainty?	2.82 ± 1.21	2.70 ± 1.06
Does it describe how each treatment works?	1.62 ± 0.92	1.78 ± 1.15
Does it describe the benefits of each treatment?	1.38 ± 0.78	1.67 ± 0.95
Does it describe the risks of each treatment?	1.38 ± 0.87	1.92 ± 1.31
Does it describe what happens if no treatment is used?	1.00 ± 0.00	1.02 ± 0.13
Does it describe how the treatment choices affect overall quality of life?	1.38 ± 0.69	1.72 ± 1.19
Is it clear that there may be more than one possible treatment choice?	2.10 ± 1.54	2.38 ± 1.64
Does it provide support for shared decision-making?	3.73 ± 1.61	3.35 ± 1.58
Based on the answers to all of the above questions, rate the overall quality of the publications as a source of information about treatment choices	2.17 ± 0.64	2.23 ± 0.81

Figure 1: A. Mean total DISCERN scores (out of 80) among each question category and overall B. Mean DISCERN scores (out of 5) with standard deviations for each DISCERN question

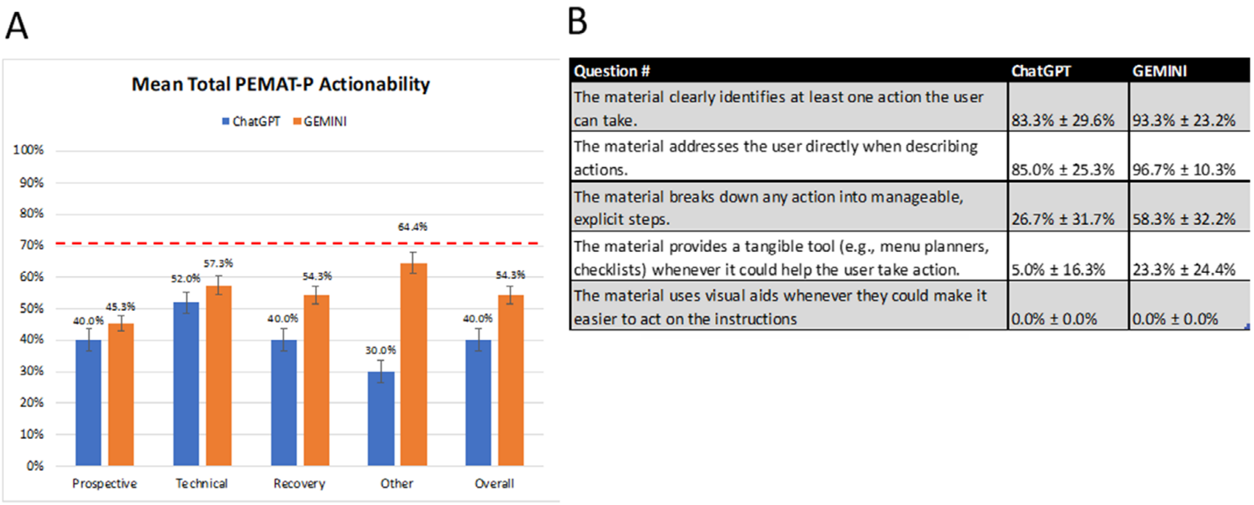
PEMAT-P understandability and actionability

The mean total PEMAT-P understandability scores for each category are represented in Figure 2. Compared to the DISCERN survey, the performance on PEMAT-P understandability was much higher. The mean total scores for responses provided by both chatbots was above 70% for each question category, indicating high understandability according to the PEMAT-P survey. Individually, 80% (16/20) of the responses provided by ChatGPT and 95% (19/20) of the responses provided by Bard were understandable. Bard scored significantly higher on “technical” questions, however, there were no significant differences among the other question categories. Overall, the majority of responses were relevant, concise, and easy to understand. However, only a handful of responses provided a summary, and no response provided any visual aid (Figure 2). The mean total PEMAT-P actionability scores for each category are represented in Figure 3. Although Bard scored slightly higher than ChatGPT, both chatbots performed very poorly. With an actionability cut-off of 70%, neither chatbot’s had any mean total scores that were above this threshold in any category. Furthermore, only one ChatGPT response and four Bard responses achieved such scores. While the vast majority of responses addressed the users directly and described at least one that could be taken, very few actionable items were provided beyond this (Figure 3).



Question #	ChatGPT	GEMINI
The material makes its purpose completely evident.	90.0% ± 24.4%	81.7% ± 20.2%
The material does not indude information or content that distracts from its purpose.	100.0% ± 0.0%	100.0% ± 0.0%
The material uses, common everyday language	96.7% ± 10.3%	100.0% ± 0.0%
Medical terms are used only to familiarize audience with the terms. When used, medical terms are defined.	91.7% ± 14.8%	93.3% ± 17.4%
The material uses the active voice.	100.0% ± 0.0%	100.0% ± 0.0%
Numbers appearing in the material are clear and easy to understand	77.5% ± 25.5%	85.0% ± 20.2%
The material does not expect the user to perform calcuations.	100.0% ± 0.0%	100.0% ± 0.0%
The material breaks or "chunks information into short sections.	83.3% ± 36.7%	96.7% ± 10.3%
The material presents information in a logical sequence.	100.0% ± 0.0%	100.0% ± 0.0%
The material provides a summary.	51.7% ± 22.9%	33.3% ± 18.7%
The material uses visual cues to draw attention to key pionts	83.3% ± 17.1%	96.7% ± 10.3%
The material uses visual aids whenever they could make content more easily understood (e.g., illustration of healthy portion size).	0.0% ± 0.0%	0.0% ± 0.0%

Figure 2: A. Mean total PEMAT-P Understandability scores among each question category and overall B. Mean PEMAT-P Understandability scores with standard deviations for each DISCERN question



Question #	ChatGPT	GEMINI
The material clearly identifies at least one action the user can take.	83.3% ± 29.6%	93.3% ± 23.2%
The material addresses the user directly when describing actions.	85.0% ± 25.3%	96.7% ± 10.3%
The material breaks down any action into manageable, explicit steps.	26.7% ± 31.7%	58.3% ± 32.2%
The material provides a tangible tool (e.g., menu planners, checklists) whenever it could help the user take action.	5.0% ± 16.3%	23.3% ± 24.4%
The material uses visual aids whenever they could make it easier to act on the instructions	0.0% ± 0.0%	0.0% ± 0.0%

Figure 3: A. Mean total PEMAT-P Actionability scores among each question category and overall B. Mean PEMAT-P Actionability scores with standard deviations for each DISCERN question

Discussion

AI chatbots have exploded in popularity in the last year. Following ChatGPT's release in December 2022, it quickly became the fastest growing internet platform in history [20]. Alarmed by ChatGPT's success, Google released Bard in March 2023 as a direct competitor [21]. Their rise in popularity has generated robust conversation in the medical community regarding their use in healthcare. With proposed functions ranging from helping with administrative duties, such as creating call schedules and handling insurance claims, to developing differential diagnoses and answering patient questions, the possible uses for AI chatbots are vast [22,23]. However, this has raised questions regarding the ethical and legal implications and limitations of their use [22,23]. To our knowledge, this is the first study to analyze and compare the quality of information provided by ChatGPT and Bard regarding laparoscopic donor nephrectomies. In our evaluation, the overall quality of responses provided by both chatbots was low, with the responses consistently failing to meet the criteria for both validated tools. The vast majority of DISCERN scores received a "poor" rating. Furthermore, only one response received a "fair" rating, while no responses received an "excellent" or even "good" rating. The responses consistently failed to provide sufficient information about the risks of kidney donation and the long-term implications on quality of life. No response provided any details about the sourcing of the information, and additional information was given only when prompted. The poor scores on the DISCERN tool represent a severe lack of quality information for potential kidney donors, which is concerning given ChatGPT and Bard's accessibility and ease of use. The performances were similarly poor on the PEMAT-P Actionability survey, with only 5 out of 40 responses achieving the 70% threshold. Aside from encouraging further discussion with a medical professional, few actionable items were provided beyond this. Because successful kidney donation requires willing participation by a healthy patient, the information provided by ChatGPT and Bard does little to help address the shortage of kidneys available for transplant in the US. In contrast, the scores on the PEMAT-Understandability were very high, with 35 out of 40 total responses achieving the 70% threshold. In general, the responses provided by both chatbots were confident, well-organized, and easy to understand. Alone, these are strengths that should be preserved and improved upon as the AI chatbot technology advances. However, this is particularly concerning in context with the poor performances on the DISCERN and PEMAT-P Actionability tools. The easy understandability of the responses could provide an illusion of high-quality and educational information. The risk of spreading "believable misinformation" to potential donors which could have significant consequences. This could be exacerbated by the fact that ChatGPT and Bard rarely provide a disclosure that they cannot give medical advice. Our findings are consistent to those reported in other studies that assessed ChatGPT (GPT-3.5 and 4.0) responses. To the questions on vaccination and immunization, ChatGPT's responses are reported to be both accurate and comprehensive for commonly asked questions, while content of the responses becomes haphazard and lacks consistency for less common questions [24]. This could reflect our finding as kidney nephrectomy being a more specialized topic than vaccination and immunization. Similarly, the study that evaluated ChatGPT's responses to 284 medical questions across 17 specialties revealed that close to 60% of the responses were almost all correct or correct while responses to the questions rated hard by the physicians were less likely to be accurate [25]. The LLM technology has received significant attention for medical education since its recent accomplishment of passing the US Medical Licensing Exam [26]. However, both the prior and our findings demonstrate that there is a need for further research and model development before the LLM technology becomes a reliable tool in the medical field, and meanwhile the technology should be used with caution, especially for specialized areas. There are several limitations to this study. First, ChatGPT and Bard are relatively new technologies that are constantly being updated, with features being changed and added. The weaknesses we discovered in this study may be addressed with future updates, demonstrating the need for ongoing evaluation. Second, little context was provided to the chatbots before entering questions. The responses provided by ChatGPT and Bard are affected by the context of the previous conversation. Had additional information been provided, the responses may have been different. Third, the scoring with the DISCERN tool, especially in the partial (2-4) range, is inherently subjective. This is reflected by the relatively poor interrater reliability for the DISCERN scores. This likely affected scoring on the DISCERN tool, as interrater reliability for PEMAT-P Understandability and Actionability was much higher. Given the growing popularity of AI chatbots, clinicians should be aware that patients who have either agreed to or are considering live kidney donation potentially received information from ChatGPT or Bard. Physicians should inquire if and how the information patients received from chatbots influenced their decision. Furthermore, clinicians should be ready to answer any questions or correct any misinformation they may have received. While ChatGPT and Bard have potential for use by patients seeking information about kidney donation, caution should be used. Two avenues in which using AI chatbots could potentially be beneficial for the kidney transplant community are a) developing domain-specific or personalized chatbots, b) utilizing AI chatbots as an introductory tool to match with specialists or mentors. For different stages of the donation process, domain-specific chatbots could be fine-tuned and tested for the sole purpose of serving kidney transplant questions. Personalized chatbots could be equipped with certain clinical information of potential donors such as lab results to generate personalized guidelines. Current chatbots could also be coupled with live specialists or mentors as an introduction to the donation process. If chatbots could be trained to match potential donors to live donor mentorship programs and furthermore select a perfect mentor match for them, depending on their personal preferences, it could potentially reduce the dropout rate of potential donors. Prospective studies must test whether using such chatbots in concordance with a mentor could increase center-level live donation rate. Currently there are no chatbots that can provide this service.

Conclusion

AI chatbots provide specific answers in response to direct questions, which are generally relevant and easy to understand by a layperson. However, the quality and actionability of the information they provide is questionable. With their rise in use, physicians should be aware of their limitations and potential for misinformation.

Author Contributions

M.W: Conceptualization (supporting), writing-original draft (lead), formal analysis (lead), data curation (equal), project administration (lead)

I.D: Data curation (equal), writing-review and editing (equal)

S.H: Methodology (supporting), data curation (equal), writing-review and editing (equal)

M.N: Writing-review and editing (equal)

N.K: Writing-review and editing (equal)

O.E. Conceptualization (lead), methodology (lead), supervision (lead), writing-review and editing (equal)

Acknowledgement

There are no acknowledgements for this research.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Dols LFC, Kok NFM, IJzermans JNM (2010) Live donor nephrectomy: A review of evidence for surgical techniques. Transpl Int 23(2):121-130.
2. Shockcor NM, Sultan S, Alvarez-Casas J, Brazio PS, Phelan M, et al. (2018) Minimally invasive donor nephrectomy: Current state of the art. Langenbecks Arch Surg 403(6):681-691.
3. Yang A, Barman N, Chin E, Herron D, Arvelakis A, et al. (2018) Robotic-assisted vs laparoscopic donor nephrectomy: A retrospective comparison of perioperative course and postoperative outcome after 1 year. J Robot Surg 12(2):343-350.
4. Windisch OL, Matter M, Pascual M, Sun P, Benerman D, et al. (2022) Robotic versus hand-assisted laparoscopic living donor nephrectomy: Comparison of two minimally invasive techniques in kidney transplantation. J Robot Surg 16(6):1471-1481.
5. Kortram K, Ijzermans JNM, Dor FJMF (2016) Perioperative events and complications in minimally invasive live donor nephrectomy: A systematic review and meta-analysis. Transplantation 100(11):2264-2275.
6. Yuan H, Liu L, Zheng S, Pu C, Wei Q, et al. (2013) The safety and efficacy of laparoscopic donor nephrectomy for renal transplantation: An updated meta-analysis. Transplant Proc 45(1):65-76.
7. Cohen AJ, Williams DS, Bohorquez H, Bruce DS, Carmody IC, et al. (2015) Robotic-assisted laparoscopic donor nephrectomy: Decreasing length of stay. Ochsner J 15(1):19-24.
8. HRSA organ procurement and transplantation network: National data.
9. Spardy J, Concepcion J, Yeager M, Andrade R, Braun H, et al. (2022) National analysis of recent trends in organ donation and transplantation in the united states: Toward optimizing care delivery and patient outcomes. Am Surg 89(12):5201-5209.
10. Matas AJ, Montgomery RA, Schold JD (2023) The organ shortage continues to be a crisis for patients with end-stage kidney disease. JAMA Surg 158(8):787.
11. Fox S, Duggan M (2013) 35% of U.S. adults have gone online to figure out a medical condition; of these, half followed up with a visit to a medical professional.

12. Organ donation: Don't let these myths confuse you.

13. ChatGPT: Optimizing language models for dialogue.

14. BARD.

15. Charnock D, Shepperd S, Needham G, Gann R (1999) DISCERN: An instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 53(2):105-111.

16. Charnock D (2004) Learning to DISCERN online: Applying an appraisal tool to health websites in a workshop setting. Health Educ Res 19(4):440-446.

17. PEMAT tool for audiovisual materials (PEMAT-P). Agency for healthcare research and quality.

18. Shoemaker SJ, Wolf MS, Brach C (2014) Development of the Patient Education Materials Assessment Tool (PEMAT): A new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 96(3):395-403.

19. Landis JR, Koch GG (1977) An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 33(2):363-374.

20. Curry D (2023) ChatGPT revenue and usage statistics.

21. What is Google BARD? Here's everything you need to know.

22. Altamimi I, Altamimi A, Alhumimidi AS, Altamimi A, Temsah MH (2023) Artificial Intelligence (AI) chatbots in medicine: A supplement, not a substitute. Cureus 15(6):e40922.

23. Adamopoulou E, Moussiades L (2020) Chatbots: History, technology, and applications. Mach Learn Appl.

24. Deiana G, Dettori M, Arghittu A, Azara A, Gabutti G, et al. (2023) Artificial intelligence and public health: Evaluating chatgpt responses to vaccination myths and misconceptions. Vaccines 11(7):1217.

25. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, et al. (2023) Assessing the accuracy and reliability of ai-generated medical responses: An evaluation of the chat-gpt model. Res Sq 3:2566942

26. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, et al. (2023) How does chatgpt perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 9:e45312.