JAMA Network Open

Original Investigation | Surgery

Comparison of Medical Research Abstracts Written by Surgical Trainees and Senior Surgeons or Generated by Large Language Models

Alexis M. Holland, MD; William R. Lorenz, MD; Jack C. Cavanagh, MA; Neil J. Smart, MD, PhD; Sullivan A. Ayuso, MD; Gregory T. Scarola, MS; Kent W. Kercher, MD; Lars N. Jorgensen, MD, PhD; Jeffrey E. Janis, MD; John P. Fischer, MD, MPH; B. Todd Heniford, MD

Abstract

IMPORTANCE Artificial intelligence (AI) has permeated academia, especially OpenAI Chat Generative Pretrained Transformer (ChatGPT), a large language model. However, little has been reported on its use in medical research.

OBJECTIVE To assess a chatbot's capability to generate and grade medical research abstracts.

DESIGN, SETTING, AND PARTICIPANTS In this cross-sectional study, ChatGPT versions 3.5 and 4.0 (referred to as chatbot 1 and chatbot 2) were coached to generate 10 abstracts by providing background literature, prompts, analyzed data for each topic, and 10 previously presented, unassociated abstracts to serve as models. The study was conducted between August 2023 and February 2024 (including data analysis).

EXPOSURE Abstract versions utilizing the same topic and data were written by a surgical trainee or a senior physician or generated by chatbot 1 and chatbot 2 for comparison. The 10 training abstracts were written by 8 surgical residents or fellows, edited by the same senior surgeon, at a high-volume hospital in the Southeastern US with an emphasis on outcomes-based research. Abstract comparison was then based on 10 abstracts written by 5 surgical trainees within the first 6 months of their research year, edited by the same senior author.

MAIN OUTCOMES AND MEASURES The primary outcome measurements were the abstract grades using 10- and 20-point scales and ranks (first to fourth). Abstract versions by chatbot 1, chatbot 2, junior residents, and the senior author were compared and judged by blinded surgeon-reviewers as well as both chatbot models. Five academic attending surgeons from Denmark, the UK, and the US, with extensive experience in surgical organizations, research, and abstract evaluation served as reviewers.

RESULTS Surgeon-reviewers were unable to differentiate between abstract versions. Each reviewer ranked an AI-generated version first at least once. Abstracts demonstrated no difference in their median (IQR) 10-point scores (resident, 7.0 [6.0-8.0]; senior author, 7.0 [6.0-8.0]; chatbot 1, 7.0 [6.0-8.0]; chatbot 2, 7.0 [6.0-8.0]; *P* = .61), 20-point scores (resident, 14.0 [12.0-7.0]; senior author, 15.0 [13.0-17.0]; chatbot 1, 14.0 [12.0-16.0]; chatbot 2, 14.0 [13.0-16.0]; *P* = .50), or rank (resident, 3.0 [1.0-4.0]; senior author, 2.0 [1.0-4.0]; chatbot 1, 3.0 [2.0-4.0]; chatbot 2, 2.0 [1.0-3.0]; *P* = .14). The abstract grades given by chatbot 1 were comparable to the surgeon-reviewers' grades. However, chatbot 2 graded more favorably than the surgeon-reviewers and chatbot 1. Median (IQR) chatbot 2-reviewer grades were higher than surgeon-reviewer grades of all 4 abstract versions (resident, 14.0 [12.0-17.0] vs 16.9 [16.0-17.5]; *P* = .02; senior author, 15.0 [13.0-17.0] vs 17.0 [16.5-18.0]; *P* = .03; chatbot 1, 14.0 [12.0-16.0] vs 17.8 [17.5-18.5]; *P* = .002; chatbot 2, 14.0 [13.0-16.0] vs 16.8 [14.5-18.0]; *P* = .04). When comparing the grades of the 2 chatbots, chatbot 2 gave higher median (IQR) grades

(continued)

Key Points

Question Can large language models generate convincing medical research abstracts?

Findings In this cross-sectional study comparing 10 medical abstracts written by surgical trainees and senior surgeons or generated by large language models, blinded expert surgeon-reviewers were asked to grade and rank these abstracts. There was no statistical difference in the grades or ranks of abstracts generated by the language model when compared with abstracts written by surgical trainees or senior surgeons.

Meaning These findings suggest that when appropriately trained with background literature, abstract formatting, primary research data, and a thorough prompt, chatbots can generate medical research abstracts that are difficult to distinguish from surgeon-scientist–written abstracts.

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Abstract (continued)

for abstracts than chatbot 1 (resident, 14.0 [13.0-15.0] vs 16.9 [16.0-17.5]; *P* = .003; senior author, 13.5 [13.0-15.5] vs 17.0 [16.5-18.0]; *P* = .004; chatbot 1, 14.5 [13.0-15.0] vs 17.8 [17.5-18.5]; *P* = .003; chatbot 2, 14.0 [13.0-15.0] vs 16.8 [14.5-18.0]; *P* = .01).

CONCLUSIONS AND RELEVANCE In this cross-sectional study, trained chatbots generated convincing medical abstracts, undifferentiable from resident or senior author drafts. Chatbot 1 graded abstracts similarly to surgeon-reviewers, while chatbot 2 was less stringent. These findings may assist surgeon-scientists in successfully implementing AI in medical research.

JAMA Network Open. 2024;7(8):e2425373. doi:10.1001/jamanetworkopen.2024.25373

Introduction

The introduction of artificial intelligence (AI) into the medical field has been both a promising and polarizing venture. Particularly, OpenAI Chat Generative Pretrained Transformer (ChatGPT; versions 3.5 and 4.0) is a new large language model, or chatbot, that has been trained from massive datasets to respond to prompts with sophisticated human-like answers.[1,2] Medical professionals agree that these large language models have opened the door for new possibilities in medicine but also Pandora's box. Arguments can be made for the benefit of AI in scientific research as well as for conflicts associated with AI in medicine.

The most common controversies associated with chatbots are the encroachment of plagiarism, biased training data, lack of creativity, and the spread of misinformation.[3] Many surgeon-scientists worry that chatbots pull from sources that cannot be given proper credit, leading to plagiarism and copyright infringement.[4,5] Although chatbots are trained on a plethora of information, there is little transparency in the data's origin.[1,6,7] As new reporting guidelines[7-9] recommend how to describe the role of AI in a project, publishers and editors grapple with the listing of chatbots as an author. Some argue that chatbots should not be listed as an author because they cannot take responsibility for what is written.[1,7,10-12] The ability of chatbots to generate novel ideas or think critically has also been questioned.[4,13-15] Of particular concern is the spread of misinformation.[4,10] Chatbots are not trained exclusively on medical texts, so there can be blatant inaccuracies (ie, hallucinations) in some of the AI responses.[2,16-18] Chatbots state this information with a false confidence that precludes inaccuracy unless scrutinized by a well-versed health care clinician.[18] Whether chatbots are endorsed by the scientific community or not, patients will inevitably use them to answer medical questions, so physicians should be invested in how to best validate the knowledge they emit.[4,15,18]

As a counterargument to these concerns, AI has several beneficial applications to the field of health care.[1,4,7,18-21] Chatbots have demonstrated the ability to translate text[4,11] and be integrated into hospital electronic medical records.[21] They have even passed the US Medical Licensing Examination steps 1 and 2, which are required by medical students to earn their degree.[22] The role of chatbots in scientific writing is being explored[23-25] with the goal of improving efficiency and productivity of surgeon-scientists.[4,6,10,14] If chatbots can be trained to assist in generating text for publication, scientists can devote more time to the complex pursuits involved in research.[1,2,4] The goal of our study was to train 2 chatbots to generate medical research abstracts and assess how these abstracts compared with resident- and senior author–written abstracts as reviewed by blinded, well-published surgeons in the field. Furthermore, we evaluated the ability of chatbots to grade and rank medical abstracts when taught with a rubric.

Methods

This cross-sectional study was performed at a tertiary care center in the Southeastern US and was determined exempt from review and the requirement of informed consent by the Carolinas Medical Center institutional review board. All abstracts utilized were written about a study previously approved by the Carolinas Medical Center institutional review board. This report follows Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline. The study was conducted between August 2023 and February 2024 (including data analysis).

Chatbot Training

OpenAI ChatGPT (versions 3.5 and 4.0; hereafter referred to as chatbot 1 and chatbot 2) was trained to generate medical abstracts based on provided abstracts as examples. The research residents and senior attending physician identified 10 abstracts[26-35] by our group from 2012 to 2022 that were presented at national meetings and published in surgical journals to serve as the training models. There was variation in the first author of each abstract, a junior trainee, but all studies had the same senior author (B.T.H.). These abstracts were inputted as examples of our group's writing style to provide few-shot learning (training an existing model by providing it examples to work from) for chatbot 1 and chatbot 2. The chatbots were prompted to note the similarities between the abstracts and confirm that they had saved our writing style. See eAppendix 1 in Supplement 1 for exact prompts.

Chatbot Testing and Writing

Ten additional abstracts[36-45] were used to investigate the chatbots' ability to generate scientific abstracts. These abstracts were written by 5 different trainees within the first 6 months of their research year at the same medical center between 2018 to 2023 to account for the novice period. Abstracts from the current year's research residents and fellows were excluded. All abstracts had the same senior author as the training abstracts (B.T.H.) and were submitted and presented at a variety of national and international conferences. Finally, these abstracts could only be included if we had access to the initial draft and final submitted version, the statistically analyzed research data, and a literature review of information concerning the topic of the abstract.

Once the chatbots were trained, we asked that it generate a scientific abstract based on the information provided. For each of the 10 abstracts, the chatbots were given the introduction and discussion of 3 relevant publications.[46-75] Text limitations prevented us from giving the chatbots the entire article. Next, we provided our prompt.[6,16] Specifically, we told the chatbots to generate text in the style of a senior surgeon-scientist with over 20 years of experience, like our senior author (B.T.H.). The analyzed real-world research data from each study was then pasted into the chat box. Finally, using the background literature, its knowledge as an experienced surgeon, and the data analysis, we asked both chatbot 1 and chatbot 2 to generate a version of each abstract in the trained writing style and in the specified format that was required by each national conference. An example prompt is available in eAppendix 1 in Supplement 1.

Abstract Comparison

Once chatbot 1 and chatbot 2 generated abstracts of each of the 10 studies, these were compared with the resident's first unedited draft and the senior surgeon's edited, submitted version of the same abstract. The 4 versions were deidentified and sent to 5 blinded surgeon-reviewers (J.E.J., L.N.J, J.P.F., N.J.S., and K.W.K.). The 5 surgeons come from academic practices in Denmark, the UK, and the US, and all have served as presidents or board members of international surgical organizations or editorial boards with extensive experience in research and abstract writing and grading. The reviewers were asked to independently score the 4 versions of the abstracts on a 10- and 20-point scale. The 10-point scale was based on a typical abstract rubric. The 20-point scale was based on the American Society of Plastic Surgeons, which entailed 4 categories: completeness, relevance, quality, and exposure (each worth 5 points). See eAppendix 2 in Supplement 1 for the rubrics. The reviewers were also asked to force rank the 4 abstract versions from first to fourth, with first being the best abstract and fourth being the worst, with no ties. They were asked to repeat these grading methods for all 10 abstracts for a total of 40 versions. Additionally, in a separate session, we tasked chatbot 1 and chatbot 2 with grading all 40 abstract versions. The chatbots were provided with the same instructions on a standard 10-point rubric with 10 being the best and a 20-point rubric broken into 4 categories: completeness, relevance, quality, and exposure. See eAppendix 3 in Supplement 1 for the prompt and rubric provided to the chatbots.

Statistical Analysis

Standard descriptive and comparison statistics were performed on the abstract versions using SAS version 9.4 (SAS Institute). The Fisher exact test was applied to compare categorical variables, and Kruskal-Wallis was utilized to compare continuous variables. All *P* values were 2-sided, and statistical significance was set at $P < .05$. We hypothesized that the chatbots would generate similarly graded and ranked abstracts as those written by surgical trainees and senior surgeons.

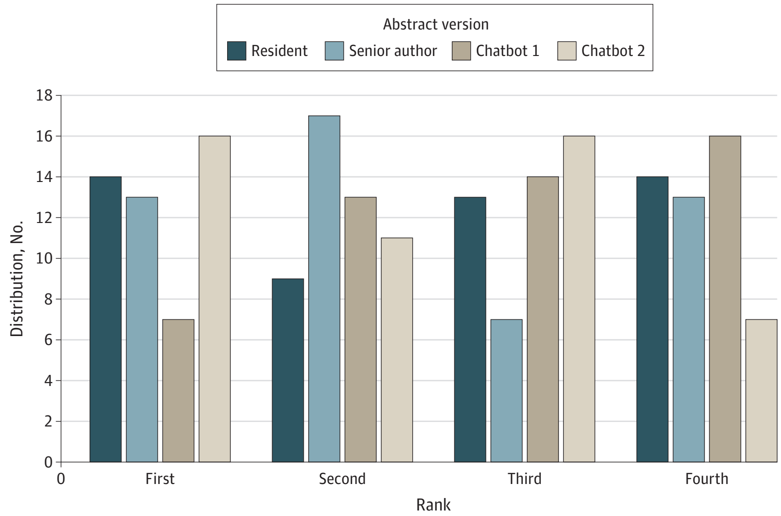
Results

Descriptive Statistics

Each surgeon-reviewer ranked an AI-generated version of an abstract first at least once, and 1 reviewer ranked either the chatbot 1 or chatbot 2 version first every time. The surgeon-reviewers ranked the resident's version first 14 of 50 times and last 14 of 50 times. They ranked the senior author's version first 13 of 50 times and last 13 of 50 times. The chatbot 1 version was ranked first least often (7 of 50 times) and ranked last most often (16 of 50 times). The chatbot 2 version was ranked first most often (16 of 50 times) and was ranked last least often (7 of 50 times) (**Figure**).

When the chatbots acted as the reviewer, chatbot 1 ranked its own version most favorably, ranking the resident's version first only 1 of 10 times, the senior author's version first 3 of 10 times, its own version first 5 of 10 times, and the chatbot 2 version first 1 of 10 times. Chatbot 1 ranked the resident's version last 3 of 10 times, the senior author's version last 2 of 10 times, its own version last 2 of 10 times, and the chatbot 2 version last 3 of 10 times. Contrastingly, chatbot 2 was more critical of its own abstracts. Chatbot 2 ranked the resident's version first 2 of 10 times and the senior author's

Figure. Distribution of Abstract Ranks by Surgeon-Reviewers



The frequency that each abstract version was ranked (first, second, third, and fourth) by surgeon-reviewers. Chatbot 1 refers to Chat Generative Pretrained Transformer (GPT) version 3.5; chatbot 2, Chat-GPT version 4.0.

version first 2 of 10 times, but it ranked the chatbot 1 version first 6 of 10 times and its own version first 0 of 10 times. Chatbot 2 never ranked chatbot 1 last and ranked itself last 4 of 10 times, the resident last 3 of 10 times, and senior author last 3 of 10 times.

When the frequency of ranks between surgeon-reviewer and chatbot-reviewer was compared, there was no statistical difference in the frequency that the resident or senior author's abstracts were ranked; however, there was a statistical difference in how the chatbot 1 version and chatbot 2 version were ranked (**Table 1**). Both the surgeon-reviewers and chatbot-reviewers ranked the resident and senior author's abstracts similarly, but they ranked chatbot 1 and chatbot 2 abstracts significantly differently. Surgeon-reviewers ranked chatbot 1 abstracts last frequently, while chatbot-reviewers did not. Surgeon-reviewers ranked chatbot 2 abstracts first frequently, while chatbot-reviewers ranked it worse.

Chatbots as Abstract Generators

There was no statistical difference in the median (IQR) 10-point scores of the resident (7.0 [6.0-8.0]), senior author (7.0 [6.0-8.0]), chatbot 1 (7.0 [6.0-8.0]), or chatbot 2 (7.0 [6.0-8.0]) (*P* = .61). Again, on the 20-point scale, the surgeon-reviewers did not prefer the resident abstracts (median [IQR] score, 14.0 [12.0-17.0]) or senior author's abstracts (median [IQR] score, 15.0 [13.0-17.0]) over the chatbot 1 (median [IQR] score, 14.0 [12.0-.16.0]) and chatbot 2 versions (median [IQR] score, 14.0 [13.0-16.0]) (*P* = .50). The reviewers' median (IQR) rank did not differ significantly between abstract versions written by residents (3.0 [1.0-4.0]) or senior authors (2.0 [1.0-4.0]) and abstract versions generated by chatbot 1 (3.0 [2.0-4.0]) or chatbot 2 (2.0 [1.0-3.0]) (*P* = .14) (**Table 2**). When only comparing the reviews of chatbot 1 and chatbot 2, there was no statistical difference in the 10-point or 20-point scores, but the surgeon-reviewers statistically ranked chatbot 2 better (median [IQR] rank for chatbot 1, 3.0 [2.0-4.0] vs chatbot 2, 2.0 [1.0-3.0]; *P* = .02) (**Table 3**).

Chatbots as a Grader

When comparing the surgeon-reviewers with chatbot 1 as a reviewer, there was no difference in their 10-point scores, 20-point scores, or ranks of any abstract version. Contrastingly, when comparing

Table 1. Frequency of Ranks by Surgeon-Reviewers Compared With Generative Language Model-Reviewers

	Grader, No. (%)			
Abstract version and rank[a]	Surgeon (n = 50)	Chatbot 1 (n = 10)	Chatbot 2 (n = 10)	*P* value[b]
Resident				
1	14 (28.0)	1 (10.0)	2 (20.0)	.67
2	9 (18.0)	4 (40.0)	1 (10.0)	
3	13 (26.0)	2 (20.0)	4 (40.0)	
4	14 (28.0)	3 (30.0)	3 (30.0)	
Senior author				
1	13 (26.0)	3 (30.0)	2 (20.0)	.76
2	17 (34.0)	2 (20.0)	2 (20.0)	
3	7 (14.0)	3 (30.0)	3 (30.0)	
4	13 (26.0)	2 (20.0)	3 (30.0)	
Chatbot 1				
1	7 (14.0)	5 (50.0)	6 (60.0)	.02
2	13 (26.0)	2 (20.0)	2 (20.0)	
3	14 (28.0)	1 (10.0)	2 (20.0)	
4	16 (32.0)	2 (20.0)	0	
Chatbot 2				
1	16 (32.0)	1 (10.0)	0	.04
2	11 (22.0)	2 (20.0)	5 (50.0)	
3	16 (32.0)	4 (40.0)	1 (10.0)	
4	7 (14.0)	3 (30.0)	4 (40.0)	

[a] Abstracts were either written by a research resident within the first 6 months of their research year, were the final submitted version edited by a senior author, or were generated by chatbot 1 (Chat Generative Pretrained Transformer [GPT] version 3.5) or chatbot 2 (Chat-GPT version 4.0).

[b] Statistical significance was *P* < .05.

the surgeon-reviewers with chatbot 2 as a reviewer, there was a statistical difference in median grades and ranks. Particularly on the 20-point scale, chatbot 2 graded higher than the surgeon-grader for the resident's abstract version (median [IQR] grade, 14.0 [12.0-17.0] vs 16.9 [16.0-17.5]; *P* = .02), the senior author's abstract version (median [IQR] grade, 15.0 [13.0-17.0] vs 17.0 [16.5-18.0]; *P* = .03), the chatbot 1 abstract version (median [IQR] grade, 14.0 [12.0-16.0] vs 17.8 [17.5-18.5]; *P* = .002), and the chatbot 2 abstract version (median [IQR] grade, 14.0 [13.0-16.0] vs 16.8 [14.5-18.0]; *P* = .04). When the reviews by chatbot 1 and chatbot 2 were compared, again chatbot 2 gave higher median (IQR) grades for all 4 abstract versions on the 20-point scale (resident, 13.5 [13.0-15.0] vs 16.9 [16.0-17.5]; *P* = .003; senior author, 13.5 [13.0-15.5] vs 17.0 [16.5-18.0]; *P* = .004; chatbot 1, 14.5 [13.0-15.0] vs 17.8 [17.5-18.5]; *P* = .003; chatbot 2, 14.0 [13.0-15.0] vs 16.8 [14.5-18.0]; *P* = .01). See **Table 4** for full analysis.

Discussion

The first aim of this cross-sectional study was to evaluate if chatbots could generate scientific abstracts as well as a research resident or senior author. Based on 10- and 20-point scales, the abstracts were not differentiable. When force ranked, the chatbot 2 version was ranked first most frequently and the chatbot 1 version was ranked last most frequently. The second goal of this study was to assess how similarly chatbot- and surgeon-reviewers could grade abstracts. Chatbot 1 abstract grades were comparable to the surgeon-reviewers' grades. However, chatbot 2 graded more favorably than the surgeon-reviewers and chatbot 1. Further observations were that the chatbots consistently utilized the provided results and did not hallucinate new data.

Although editors have worked quickly to regulate the implementation of AI in scientific writing, if it is permitted at all,[14] AI continues to permeate all fields of medicine, academia, and research.[1,4] The goal of this study was to evaluate if chatbots could generate and grade medical research abstracts. We found that, when trained using real-world data, chatbots could generate medical research abstracts in a manner that was not able to be differentiated from a human researcher. This is a promising and exciting observation, but further exploration should elucidate the ability of chatbots to consistently grade abstracts, given that the ability varied between chatbots 1 and 2 in our study. There are a variety of rubrics and scoring systems utilized in consideration for national meetings, but our findings indicate that a greater range point-system with defined categories is helpful to discern abstract quality. Abstract grading and consideration are time consuming, but the chatbots showed the potential to expedite this process and could help narrow down the number of abstracts human-reviewers need to read. Our group continues to explore the capability of chatbots as an abstract grader by more extensively training the AI model.

Table 2. Chatbots an Abstract Generator: Comparison of Grades by Surgeon-Reviewers[a]

	Grade by surgeon reviewer, median (IQR)				
Grading scale	Resident	Senior author	Chatbot 1	Chatbot 2	*P* value[b]
10-Point scale	7.0 (6.0-8.0)	7.0 (6.0-8.0)	7.0 (6.0-8.0)	7.0 (6.0-8.0)	.61
20-Point scale	14.0 (12.0-17.0)	15.0 (13.0-17.0)	14.0 (12.0-16.0)	14.0 (13.0-16.0)	.50
Rank	3.0 (1.0-4.0)	2.0 (1.0-4.0)	3.0 (2.0-4.0)	2.0 (1.0-3.0)	.14

[a] Abstracts were either written by a research resident within the first 6 months of their research year, were the final submitted version edited by a senior author, or were generated by chatbot 1 (Chat Generative Pretrained Transformer [GPT] version 3.5) or chatbot 2 (Chat-GPT version 4.0).

[b] Statistical significance was *P* < .05.

Table 3. Chatbots as an Abstract Generator: Comparison of Grades Subgroup Analysis: Chatbot 1 vs Chatbot 2[a]

	Grade by surgeon reviewer, median (IQR)		
Grading scale	Chatbot 1	Chatbot 2	*P* value[b]
10-Point scale	7.0 (6.0-8.0)	7.0 (6.0-8.0)	.41
20-Point scale	14.0 (12.0-16.0)	14.0 (13.0-16.0)	.41
Rank	3.0 (2.0-4.0)	2.0 (1.0-3.0)	.02

[a] Abstracts were generated by chatbot 1 (Chat Generative Pretrained Transformer [GPT] version 3.5) or chatbot 2 (Chat-GPT version 4.0) and graded by 5 surgeon-reviewers.

[b] Statistical significance was *P* < .05.

Despite successful implementation of AI in numerous areas of academia, like all new technologies, there is hesitancy to change.[15] Chatbots gather information from unknown sources that cannot be directly cited, leading to controversy over plagiarism and copyright infringement.[4,5] To combat this ethical dilemma, some investigators have asked chatbots to provide a list of references,[4,13] but when cross-checked, the sources chatbots provided were sometimes falsified.[10,11] In the medical field, where patient privacy is extremely important, there is a particular worry about the security of patient information shared with chatbots.[1,4] Detractors have labeled chatbots a "stochastic parrot"[1] that "threatens the trajectory"[13] of modern medicine and scientific research. Some believe chatbots will stifle creativity, replace the learned ability of students to write papers, and degrade the sense of academic integrity.[14,15,76] The counterargument is that learners still develop these writing skills, but in a nontraditional way, by editing chatbot output.[15]

Arguably, the most pertinent debate against chatbots is the spread of misinformation.[4,10] The hallucinations[18] produced by chatbots may present as fake statistics[77] or inaccurate answers to medical questions. Emile et al[78] assessed a chatbot's ability to answer common questions about colon cancer, and Samaan et al[79] reviewed the accuracy of a chatbot's answers regarding bariatric surgery. Both found that the responses were mostly accurate, but there were certainly incorrect answers as well.[78,79] Patients using chatbots may not be able to discern fact from fiction, so physicians, whether they support AI or not, should be invested in how their patients are using it.[4,15,18]

Despite these concerns, chatbots have potential in the medical community, including the potential to boost productivity in scientific writing. Chatbots can save researchers time by formatting papers specific to a journal,[1,4] running statistics,[18] and accelerating the publishing process, which alleviates pressure on surgeon-scientists.[4,6,10,14] Chatbots can also be leveraged to reduce effort spent preparing a manuscript or grant by editing preexisting text, enhancing readability, and decreasing the number of rounds of feedback between authors.[4,10] By increasing efficiency, some believe that chatbots can provide time to devote to more valuable pursuits.[1,2,4] The ultimate goal of medical research is to advance knowledge and improve health for patients, so if we can employ AI[1,4] to perform the routine tasks of research, we can spend more time on the creative aspects, complex questions, and critical thinking involved in research.

Table 4. Chatbots as a Grader: Comparison of Grades and Ranks Given by Surgeon-Reviewers vs Chatbot-Reviewers[a]

	Grade, median (IQR)			Grade, median (IQR)			Grade, median (IQR)		
Abstract version	Surgeon-grader	Chatbot 1-grader	*P* value[b]	Surgeon-grader	Chatbot 2-grader	*P* value[b]	Chatbot 1-grader	Chatbot 2-grader	*P* value[b]
10-Point scale									
Resident	7.0 (6.0-8.0)	7.0 (6.7-7.5)	.89	7.0 (6.0-8.0)	7.5 (7.5-7.8)	.24	7.0 (6.7-7.5)	7.5 (7.5-7.8)	.12
Senior author	7.0 (6.0-8.0)	7.3 (6.4-8.0)	.86	7.0 (6.0-8.0)	7.5 (7.5-7.8)	.13	7.3 (6.4-8.0)	7.5 (7.5-7.8)	.30
Chatbot 1	7.0 (6.0-8.0)	7.2 (6.5-7.8)	.10	7.0 (6.0-8.0)	8.2 (8.0-8.5)	.003	7.2 (6.5-7.8)	8.2 (8.0-8.5)	.02
Chatbot 2	7.0 (6.0-8.0)	7.3 (6.2-7.5)	.76	7.0 (6.0-8.0)	7.9 (7.0-8.0)	.14	7.3 (6.2-7.5)	7.9 (7.0-8.0)	.08
20-Point scale									
Resident	14.0 (12.0-17.0)	14.0 (13.0-15.0)	.79	14.0 (12.0-17.0)	16.9 (16.0-17.5)	.02	14.0 (13.0-15.0)	16.9 (16.0-17.5)	.003
Senior author	15.0 (13.0-17.0)	13.5 (13.0-15.5)	.28	15.0 (13.0-17.0)	17.0 (16.5-18.0)	.03	13.5 (13.0-15.5)	17.0 (16.5-18.0)	.004
Chatbot 1	14.0 (12.0-16.0)	14.5 (13.0-15.0)	.48	14.0 (12.0-16.0)	17.8 (17.5-18.5)	.002	14.5 (13.0-15.0)	17.8 (17.5-18.5)	.003
Chatbot 2	14.0 (13.0-16.0)	14.0 (13.0-15.0)	.79	14.0 (13.0-16.0)	16.8 (14.5-18.0)	.04	14.0 (13.0-15.0)	16.8 (14.5-18.0)	.01
Rank, quartile (range)									
Resident	3.0 (1.0-4.0)	2.5 (2.0-4.0)	.70	3.0 (1.0-4.0)	3.0 (2.0-4.0)	.54	2.5 (2.0-4.0)	3.0 (2.0-4.0)	.78
Senior author	2.0 (1.0-4.0)	2.5 (1.0-3.0)	>.99	2.0 (1.0-4.0)	3.0 (2.0-4.0)	.45	2.5 (1.0-3.0)	3.0 (2.0-4.0)	.56
Chatbot 1	3.0 (2.0-4.0)	1.5 (1.0-3.0)	.05	3.0 (2.0-4.0)	1.0 (1.0-2.0)	.002	1.5 (1.0-3.0)	1.0 (1.0-2.0)	.51
Chatbot 2	2.0 (1.0-3.0)	3.0 (2.0-4.0)	.10	2.0 (1.0-3.0)	2.5 (2.0-4.0)	.11	3.0 (2.0-4.0)	2.5 (2.0-4.0)	.94

[a] Abstracts were either written by a research resident within the first 6 months of their research year, were the final submitted version edited by a senior author, or were generated by chatbot 1 (Chat Generative Pretrained Transformer [GPT] version 3.5) or chatbot 2 (Chat-GPT version 4.0).

[b] Statistical significance was $P < .05$.

Prior studies have investigated the ability of chatbots to regenerate available medical research abstracts. Gao et al[77] provided a chatbot with the title and journal name of previously published abstracts, while Levin et al[23] provided the title and results section and asked it to regenerate the text. Gao et al[77] found that human-reviewers correctly identified 68% of the chatbot-written abstracts and 86% of the human-written abstracts, but the chatbot versions were noted to be vague, making it easier to correctly distinguish them. Levin et al[23] showed that AI-generated versions had fewer grammatical errors and more unique words than the scientist-written version, making these more difficult to distinguish.[24,25]

This study stands apart from prior work on AI-writing because the chatbots were provided with more than just a title and journal name.[77] By training chatbots to generate text in our group's writing style and inputting background, previously published studies, and statistically analyzed data for each abstract, we combatted the tendency for chatbots to hallucinate results. We suspect that as chatbots become more sophisticated, the potential to generate abstracts may surpass the ability of some researchers and may expand to generating full manuscripts.

One of the interesting observations we encountered while working with the chatbots was the variation between the chatbots 1 and 2. Both chatbot 1 and chatbot 2 were trained with data extending until September 2021, but chatbot 2 is considered the more advanced version[80] and in our experience, had more independent thinking.[81] When asking the chatbots to generate text, we used the same online session to provide consistency. Chatbot 1 was compliant and completed the tasks without needing redirection, but chatbot 2 had difficulty complying, required restarting new sessions, retraining each one, and several reminders of the prompt to finish writing all 10 abstracts. Although we intended to train the chatbots on more than 10 abstracts, often after the fifth abstract, chatbot 2 pushed back, stating that it did not need more abstracts to learn the writing style. We proceeded, however, in training the chatbots with 10 abstracts. Despite chatbot 2 being less compliant, blinded surgeons agreed that the chatbot 2 abstract versions were better and more consistent than the chatbot 1 versions. The chatbots followed directions on grading more easily, suggesting future promise in saving researchers and editors' time.

Both advocates and skeptics mostly agree that chatbots will not replace surgeons as primary decision makers in the near future.[4,6,17,21] AI has the potential to complement patient-clinician interactions and assist in medical research, but it will be difficult for AI to replace a surgeon's judgement.[6,17,21] Chatbots can serve as a helpful ally in medical abstract generating and grading, but at this point in its evolution, AI cannot perform independently. In the meantime, our goal is to leverage AI for the function of better research and ultimately better patient care.[4,14] AI is permeating all facets of medicine, and as clinicians, we need to decide the best approach to incorporate it into our research and clinical space.

Limitations

The primary limitation of this study was the small sample size of abstracts and reviewers. To combat this limitation, we intentionally chose surgeons who had extensive experience and represented different practice models and international backgrounds. Furthermore, this work is based on abdominal wall reconstruction abstracts and thus may not translate to other fields of medicine. There are also limitations of chatbots. The chatbots have a knowledge cutoff in September 2021 and do not have the ability to browse the internet for more recent context. Chatbots are dependent on the data and training they received, which could result in bias that they learned.[3,82] The chatbots additionally have a token cutoff, or character limit, which may inhibit the quantity of training or prompting the model can learn at a time.[17]

Conclusions

The findings of this cross-sectional study suggest that a chatbot can generate quality medical research abstracts when the user spends the time to train it, feed it background information, and supply it with analyzed data. The chatbots in this study also demonstrated the ability to grade abstracts, with chatbot 2 being less stringent than chatbot 1. The findings of this study serve as an example of successful and safe implementation of AI in scientific writing, which we hope is considered as editors and publishers continue to determine the regulation and acceptable role of AI.

ARTICLE INFORMATION

Accepted for Publication: June 4, 2024.

Published: August 2, 2024. doi:10.1001/jamanetworkopen.2024.25373

Corresponding Author: B. Todd Heniford, MD, Division of Gastrointestinal and Minimally Invasive Surgery, Department of Surgery, Atrium Health Carolinas Medical Center, 1025 Morehead Medical Dr, Ste 300, Charlotte, NC 28204 (todd.heniford@gmail.com).

Author Affiliations: Division of Gastrointestinal and Minimally Invasive Surgery, Department of Surgery, Atrium Health Carolinas Medical Center, Charlotte, North Carolina (Holland, Lorenz, Ayuso, Scarola, Kercher, Heniford); Department of Economics, Massachusetts Institute of Technology, Cambridge (Cavanagh); Division of Colorectal Surgery, Department of Surgery, Royal Devon & Exeter Hospital, Exeter, Devon, United Kingdom (Smart); Department of Clinical Medicine, University of Copenhagen, Bispedjerg & Frederiksberg Hospital, Copenhagen, Denmark (Jorgensen); Division of Plastic and Reconstructive Surgery, The Ohio State University Wexner Medical Center, Columbus (Janis); Division of Plastic Surgery, University of Pennsylvania Health System, Philadelphia (Fischer).

Author Contributions: Dr Heniford had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Holland, Smart, Ayuso, Scarola, Janis, Fischer, Heniford.

Acquisition, analysis, or interpretation of data: Holland, Lorenz, Cavanagh, Smart, Ayuso, Scarola, Kercher, Jorgensen, Heniford.

Drafting of the manuscript: Holland, Ayuso, Scarola, Kercher, Janis, Fischer.

Critical review of the manuscript for important intellectual content: Holland, Lorenz, Cavanagh, Smart, Ayuso, Scarola, Kercher, Jorgensen, Janis, Heniford.

Statistical analysis: Holland, Scarola.

Administrative, technical, or material support: Lorenz, Cavanagh, Ayuso, Scarola, Kercher, Janis, Heniford.

Supervision: Smart, Ayuso, Scarola, Kercher, Janis, Fischer, Heniford.

Conflict of Interest Disclosures: Dr Smart reported receiving personal fees from Medtronic and Gore outside the submitted work. Dr Heniford reported receiving grants from Gore outside the submitted work. No other disclosures were reported.

Funding/Support: This study was funded by Atrium Health and Carolinas Laparoscopic and Advanced Surgery Program.

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Disclaimer: Neither artificial intelligence nor large language models were employed in the writing of this manuscript.

Data Sharing Statement: See Supplement 2.

REFERENCES

1. Lund BD, Wang T, Mannuru NR, Nie B, Shimray S, Wang Z. ChatGPT and a new academic reality: artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing. *J Assoc Inf Sci Technol*. 2023;74(5):570-581. doi:10.1002/asi.24750

2. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature*. 2023; 614(7947):224-226. doi:10.1038/d41586-023-00288-7

3. Acerbi A, Stubbersfield JM. Large language models show human-like content biases in transmission chain experiments. *Proc Natl Acad Sci U S A*. 2023;120(44):e2313790120. doi:10.1073/pnas.2313790120

4. Borger JG, Ng AP, Anderton H, et al. Artificial intelligence takes center stage: exploring the capabilities and implications of ChatGPT and other AI-assisted technologies in scientific research and education. *Immunol Cell Biol*. 2023;101(10):923-935. doi:10.1111/imcb.12689

5. Dehouche N. Plagiarism in the age of massive generative pre-trained transformers (GPT-3). *Ethics Sci Environ Polit*. 2021;21:17-23. doi:10.3354/esep00195

6. Gupta R, Park JB, Bisht C, et al. Expanding cosmetic plastic surgery research with ChatGPT. *Aesthet Surg J*. 2023;43(8):930-937. doi:10.1093/asj/sjad069

7. Ibrahim H, Liu X, Denniston AK. Reporting guidelines for artificial intelligence in healthcare research. *Clin Exp Ophthalmol*. 2021;49(5):470-476. doi:10.1111/ceo.13943

8. Vasey B, Nagendran M, Campbell B, et al; DECIDE-AI expert group. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ*. 2022;377: e070904. doi:10.1136/bmj-2022-070904

9. Ibrahim H, Liu X, Rivera SC, et al. Reporting guidelines for clinical trials of artificial intelligence interventions: the SPIRIT-AI and CONSORT-AI guidelines. *Trials*. 2021;22(1):11. doi:10.1186/s13063-020-04951-6

10. Rafaqat W, Chu DI, Kaafarani HMAI. AI and ChatGPT meet surgery: a word of caution for surgeon-scientists. *Ann Surg*. 2023;278(5):e943-e944. doi:10.1097/SLA.0000000000006000

11. Kim SG. Using ChatGPT for language editing in scientific articles. *Maxillofac Plast Reconstr Surg*. 2023;45(1):13. doi:10.1186/s40902-023-00381-x

12. ElHawary H, Gorgy A, Janis JE. Large language models in academic plastic surgery: the way forward. *Plast Reconstr Surg Glob Open*. 2023;11(4):e4949. doi:10.1097/GOX.0000000000004949

13. Seth I, Bulloch G, Lee CHA. Redefining academic integrity, authorship, and innovation: the impact of ChatGPT on surgical research. *Ann Surg Oncol*. 2023;30(8):5284-5285. doi:10.1245/s10434-023-13642-w

14. Tel A, Parodi PC, Robiony M, Zanotti B, Zingaretti N. Letter to the editor: could ChatGPT improve knowledge in surgery? *Ann Surg Oncol*. 2023;30(7):3942-3943. doi:10.1245/s10434-023-13518-z

15. Doyal AS, Sender D, Nanda M, Serrano RA. ChatGPT and artificial intelligence in medical writing: concerns and ethical considerations. *Cureus*. 2023;15(8):e43292.

16. Chen Q, Sun H, Liu H, et al. An extensive benchmark study on biomedical text generation and mining with ChatGPT. *Bioinformatics*. 2023;39(9):btad557. doi:10.1093/bioinformatics/btad557

17. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. *JAMA*. 2023; 330(9):866-869. doi:10.1001/jama.2023.14217

18. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. *N Engl J Med*. 2023; 388(13):1233-1239. doi:10.1056/NEJMsr2214184

19. Ayuso SA, Elhage SA, Zhang Y, et al. Predicting rare outcomes in abdominal wall reconstruction using image-based deep learning models. *Surgery*. 2023;173(3):748-755. doi:10.1016/j.surg.2022.06.048

20. Elhage SA, Deerenberg EB, Ayuso SA, et al. Development and validation of image-based deep learning models to predict surgical complexity and complications in abdominal wall reconstruction. *JAMA Surg*. 2021;156(10): 933-940. doi:10.1001/jamasurg.2021.3012

21. El Hechi M, Ward TM, An GC, et al. Artificial intelligence, machine learning, and surgical science: reality versus hype. *J Surg Res*. 2021;264:A1-A9. doi:10.1016/j.jss.2021.01.046

22. Gilson A, Safranek CW, Huang T, et al. 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*. 2023;9:e45312. doi:10.2196/45312

23. Levin G, Meyer R, Kadoch E, Brezinov Y. Identifying ChatGPT-written OBGYN abstracts using a simple tool. *Am J Obstet Gynecol MFM*. 2023;5(6):100936. doi:10.1016/j.ajogmf.2023.100936

24. Levin G, Pareja R, Viveros-Carreño D, et al. Association of reviewer experience with discriminating human-written versus ChatGPT-written abstracts. *Int J Gynecol Cancer*. 2024;34(5):669-674. doi:10.1136/ijgc-2023-005162

25. Levin G, Meyer R, Yasmeen A, et al. Chat generative pre-trained transformer-written obstetrics and gynecology abstracts fool practitioners. *Am J Obstet Gynecol MFM*. 2023;5(8):100993. doi:10.1016/j.ajogmf.2023.100993

26. Elhage SA, Deerenberg EB, Ayuso SA, et al. Development and validation of image-based deep learning models to predict surgical complexity and complications in abdominal wall reconstruction. *JAMA Surg*. 2021;156(10): 933-940. doi:10.1001/jamasurg.2021.3012

27. Shao JM, Deerenberg EB, Elhage SA, et al. Are laparoscopic and open ventral hernia repairs truly comparable?: a propensity-matched study in large ventral hernias. *Surg Endosc*. 2021;35(8):4653-4660. doi:10.1007/s00464-020-07894-w

28. Ross SW, Wormer BA, Kim M, et al. Defining surgical outcomes and quality of life in massive ventral hernia repair: an international multicenter prospective study. *Am J Surg*. 2015;210(5):801-813. doi:10.1016/j.amjsurg.2015.06.020

29. Schlosser KA, Maloney SR, Prasad T, Colavita PD, Augenstein VA, Heniford BT. Three-dimensional hernia analysis: the impact of size on surgical outcomes. *Surg Endosc*. 2020;34(4):1795-1801. doi:10.1007/s00464-019-06931-7

30. Shao JM, Deerenberg EB, Elhage SA, et al. Recurrent incisional hernia repairs at a tertiary hernia center: are outcomes really inferior to initial repairs? *Surgery*. 2021;169(3):580-585. doi:10.1016/j.surg.2020.10.009

31. Schlosser KA, Maloney SR, Prasad T, Colavita PD, Augenstein VA, Heniford BT. Too big to breathe: predictors of respiratory failure and insufficiency after open ventral hernia repair. *Surg Endosc*. 2020;34(9):4131-4139. doi:10.1007/s00464-019-07181-3

32. Ayuso SA, Elhage SA, Zhang Y, et al. Predicting rare outcomes in abdominal wall reconstruction using image-based deep learning models. *Surgery*. 2023;173(3):748-755. doi:10.1016/j.surg.2022.06.048

33. Colavita PD, Tsirline VB, Belyansky I, et al. Prospective, long-term comparison of quality of life in laparoscopic versus open ventral hernia repair. *Ann Surg*. 2012;256(5):714-722. doi:10.1097/SLA.0b013e3182734130

34. Sacco JM, Ayuso SA, Salvino MJ, et al. Preservation of deep epigastric perforators during anterior component separation technique (ACST) results in equivalent wound complications compared to transversus abdominis release (TAR). *Hernia*. 2023;27(4):819-827. doi:10.1007/s10029-023-02811-1

35. Deerenberg EB, Shao JM, Elhage SA, et al. Preoperative botulinum toxin A injection in complex abdominal wall reconstruction- a propensity-scored matched study. *Am J Surg*. 2021;222(3):638-642. doi:10.1016/j.amjsurg.2021.01.010

36. Ayuso SA, Aladegbami BG, Kercher KW, Colavita PD, Augenstein VA, Heniford BT. Coated Polypropylene Mesh Is Associated With Increased Infection in Abdominal Wall Reconstruction. *J Surg Res*. 2022;275:56-62. doi:10.1016/j.jss.2022.01.027

37. Kao AM, Huntington CR, Otero J, et al. Emergent Laparoscopic Ventral Hernia Repairs. *J Surg Res*. 2018;232:497-502. doi:10.1016/j.jss.2018.07.034

38. Wilson H, Tawkaliyar R, Rose M, et al OC-028 Using The Vacuum Assisted "French Fry" Technique (Fft) For Wound Closure In Contaminated Open Abdominal Wall Reconstruction (AWR). *Br J Surg*. 2023;110(Suppl 2):znad080.035. doi:10.1093/bjs/znad080.035

39. Wilson HH, Ma C, Ku D, et al. procedure volume impacts complications and length of stay following emergent paraesophageal hernia repair. Abstract presented at: 2023 Session of the Society of American Gastrointestinal and Endoscopic Surgeons; 2023; Montreal, Canada. Accessed July 15, 2024. https://link.springer.com/article/10.1007/s00464-023-10072-3

40. Wilson HH, Ayuso SA, Rose M, et al. Defining surgical risk in octogenarians undergoing paraesophageal hernia repair. *Surg Endosc*. 2023;37(11):8644-8654. doi:10.1007/s00464-023-10270-z

41. Elhage SA, Ayuso SA, Deerenberg EB, et al. Factors Predicting Increased Length of Stay in Abdominal Wall Reconstruction. *Am Surg*. 2023;89(5):1539-1545. doi:10.1177/00031348211047503

42. Wilson HH, Ku D, Scarola GT, Kearns J, Heniford BT. A Pilot study evaluating the influence of disparities in prostate cancer at diagnosis. Abstract presented at: North Carolina/South Carolina American College of Surgeons 2022 Annual Meeting; October 2022; Myrtle Beach, SC.

43. Ayuso SA, Elhage SA, Aladegbami BG, et al. Posterior component separation versus transversus abdominis release: an evaluation of wound morbidity and other perioperative outcomes. Abstract presented at the 45th European Hernia Society Annual International Congress; May 2023; Barcelona, Spain.

44. Katzen MM, Colavita PD, Sacco JM, et al. Observational study of complex abdominal wall reconstruction using porcine dermal matrix: How have outcomes changed over 14 years? *Surgery*. 2023;173(3):724-731. doi:10.1016/j.surg.2022.08.041

45. Katzen M, Sacco J, Ku D, et al. Impact of race and ethnicity on rates of emergent ventral hernia repair (VHR): has anything changed? *Surg Endosc*. 2023;37(7):5561-5569. doi:10.1007/s00464-022-09732-7

46. van't Riet M, Burger JWA, Bonthuis F, Jeekel J, Bonjer HJ. Prevention of adhesion formation to polypropylene mesh by collagen coating: a randomized controlled study in a rat model of ventral hernia repair. *Surg Endosc*. 2004;18(4):681-685. doi:10.1007/s00464-003-9054-4

47. Deeken CR, Faucher KM, Matthews BD. A review of the composition, characteristics, and effectiveness of barrier mesh prostheses utilized for laparoscopic ventral hernia repair. *Surg Endosc*. 2012;26(2):566-575. doi:10.1007/s00464-011-1899-3

48. Thomas JD, Fafaj A, Zolin SJ, et al. Non-coated versus coated mesh for retrorectus ventral hernia repair: a propensity score-matched analysis of the Americas Hernia Society Quality Collaborative (AHSQC). *Hernia*. 2021;25(3):665-672. doi:10.1007/s10029-020-02229-z

49. Olmi S, Cesana G, Erba L, Croce E. Emergency laparoscopic treatment of acute incarcerated incisional hernia. *Hernia*. 2009;13(6):605-608. doi:10.1007/s10029-009-0525-y

50. MacDonald E, Pringle K, Ahmed I. Single port laparoscopic repair of incarcerated ventral hernia. Re: Laparoscopic repair of incarcerated ventral abdominal wall hernias, Shah RH et al. (2008) Hernia 12(5):457-463. *Hernia*. 2009;13(3):339-339. doi:10.1007/s10029-009-0492-3

51. Helgstrand F, Rosenberg J, Kehlet H, Bisgaard T. Outcomes after emergency versus elective ventral hernia repair: a prospective nationwide study. *World J Surg*. 2013;37(10):2273-2279. doi:10.1007/s00268-013-2123-5

52. Ayuso SA, Elhage SA, Aladegbami BG, et al. Delayed primary closure (DPC) of the skin and subcutaneous tissues following complex, contaminated abdominal wall reconstruction (AWR): a propensity-matched study. *Surg Endosc*. 2022;36(3):2169-2177. doi:10.1007/s00464-021-08485-z

53. Soares KC, Baltodano PA, Hicks CW, et al. Novel wound management system reduction of surgical site morbidity after ventral hernia repairs: a critical analysis. *Am J Surg*. 2015;209(2):324-332. doi:10.1016/j.amjsurg.2014.06.022

54. Berner-Hansen V, Oma E, Willaume M, Jensen KK. Prophylactic negative pressure wound therapy after open ventral hernia repair: a systematic review and meta-analysis. *Hernia*. 2021;25(6):1481-1490. doi:10.1007/s10029-021-02485-7

55. Chimukangara M, Helm MC, Frelich MJ, et al. A 5-item frailty index based on NSQIP data correlates with outcomes following paraesophageal hernia repair. *Surg Endosc*. 2017;31(6):2509-2519. doi:10.1007/s00464-016-5253-7

56. Hosein S, Carlson T, Flores L, Armijo PR, Oleynikov D. Minimally invasive approach to hiatal hernia repair is superior to open, even in the emergent setting: a large national database analysis. *Surg Endosc*. 2021;35(1):423-428. doi:10.1007/s00464-020-07404-y

57. Sherrill W III, Rossi I, Genz M, Matthews BD, Reinke CE. Non-elective paraesophageal hernia repair: surgical approaches and short-term outcomes. *Surg Endosc*. 2021;35(7):3405-3411. doi:10.1007/s00464-020-07782-3

58. Poulose BK, Gosen C, Marks JM, et al. Inpatient mortality analysis of paraesophageal hernia repair in octogenarians. *J Gastrointest Surg*. 2008;12(11):1888-1892. doi:10.1007/s11605-008-0625-5

59. Sorial RK, Ali M, Kaneva P, et al. Modern era surgical outcomes of elective and emergency giant paraesophageal hernia repair at a high-volume referral center. *Surg Endosc*. 2020;34(1):284-289. doi:10.1007/s00464-019-06764-4

60. Schlottmann F, Strassle PD, Allaix ME, Patti MG. Paraesophageal Hernia Repair in the USA: Trends of Utilization Stratified by Surgical Volume and Consequent Impact on Perioperative Outcomes. *J Gastrointest Surg*. 2017;21(8):1199-1205. doi:10.1007/s11605-017-3469-z

61. Majumder A, Fayezizadeh M, Neupane R, Elliott HL, Novitsky YW. Benefits of Multimodal Enhanced Recovery Pathway in Patients Undergoing Open Ventral Hernia Repair. *J Am Coll Surg*. 2016;222(6):1106-1115. doi:10.1016/j.jamcollsurg.2016.02.015

62. Joseph WJ, Cuccolo NG, Baron ME, Chow I, Beers EH. Frailty predicts morbidity, complications, and mortality in patients undergoing complex abdominal wall reconstruction. *Hernia*. 2020;24(2):235-243. doi:10.1007/s10029-019-02047-y

63. Ueland W, Walsh-Blackmore S, Nisiewicz M, et al. The contribution of specific enhanced recovery after surgery (ERAS) protocol elements to reduced length of hospital stay after ventral hernia repair. *Surg Endosc*. 2020;34(10):4638-4644. doi:10.1007/s00464-019-07233-8

64. Iyengar S, Hall IJ, Sabatino SA. Racial/Ethnic Disparities in Prostate Cancer Incidence, Distant Stage Diagnosis, and Mortality by U.S. Census Region and Age Group, 2012-2015. *Cancer Epidemiol Biomarkers Prev*. 2020;29(7):1357-1364. doi:10.1158/1055-9965.EPI-19-1344

65. Orom H, Biddle C, Underwood W III, Homish GG, Olsson CA. Racial or Ethnic and Socioeconomic Disparities in Prostate Cancer Survivors' Prostate-specific Quality of Life. *Urology*. 2018;112:132-137. doi:10.1016/j.urology.2017.08.014

66. Lynch SM, Sorice K, Tagai EK, Handorf EA. Use of empiric methods to inform prostate cancer health disparities: Comparison of neighborhood-wide association study "hits" in black and white men. *Cancer*. 2020;126 (9):1949-1957. doi:10.1002/cncr.32734

67. Shao JM, Alimi Y, Conroy D, Bhanot P. Outcomes using indocyanine green angiography with perforator-sparing component separation technique for abdominal wall reconstruction. *Surg Endosc*. 2020;34(5):2227-2236. doi:10.1007/s00464-019-07012-5

68. Elhage SA, Marturano MN, Prasad T, et al. Impact of perforator sparing on anterior component separation outcomes in open abdominal wall reconstruction. *Surg Endosc*. 2021;35(8):4624-4631. doi:10.1007/s00464-020-07888-8

69. Maloney SR, Schlosser KA, Prasad T, et al. The impact of component separation technique versus no component separation technique on complications and quality of life in the repair of large ventral hernias. *Surg Endosc*. 2020;34(2):981-987. doi:10.1007/s00464-019-06892-x

70. Samson DJ, Gachabayov M, Latifi R. Biologic Mesh in Surgery: A Comprehensive Review and Meta-Analysis of Selected Outcomes in 51 Studies and 6079 Patients. *World J Surg*. 2021;45(12):3524-3540. doi:10.1007/s00268-020-05887-3

71. Katzen M, Ayuso SA, Sacco J, et al. Outcomes of biologic versus synthetic mesh in CDC class 3 and 4 open abdominal wall reconstruction. *Surg Endosc*. 2023;37(4):3073-3083. doi:10.1007/s00464-022-09486-2

72. Kao AM, Arnold MR, Augenstein VA, Heniford BT. Prevention and Treatment Strategies for Mesh Infection in Abdominal Wall Reconstruction. *Plast Reconstr Surg*. 2018;142(3)(suppl 3):149S-155S. doi:10.1097/PRS.0000000000004871

73. Poulose BK, Shelton J, Phillips S, et al. Epidemiology and cost of ventral hernia repair: making the case for hernia research. *Hernia*. 2012;16(2):179-183. doi:10.1007/s10029-011-0879-9

74. Katzen M, Sacco J, Ku D, et al. Impact of race and ethnicity on rates of emergent ventral hernia repair (VHR): has anything changed? *Surg Endosc*. 2023;37(7):5561-5569. doi:10.1007/s00464-022-09732-7

75. Colavita PD, Tsirline VB, Walters AL, Lincourt AE, Belyansky I, Heniford BT. Laparoscopic versus open hernia repair: outcomes and sociodemographic utilization results from the nationwide inpatient sample. *Surg Endosc*. 2013;27(1):109-117. doi:10.1007/s00464-012-2432-z

76. Gordijn B, Have HT. ChatGPT: evolution or revolution? *Med Health Care Philos*. 2023;26(1):1-2. doi:10.1007/s11019-023-10136-0

77. Gao CA, Howard FM, Markov NS, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digit Med*. 2023;6(1):75. doi:10.1038/s41746-023-00819-6

78. Emile SH, Horesh N, Freund M, et al. How appropriate are answers of online chat-based artificial intelligence (ChatGPT) to common questions on colon cancer? *Surgery*. 2023;174(5):1273-1275. doi:10.1016/j.surg.2023.06.005

79. Samaan JS, Yeo YH, Rajeev N, et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. *Obes Surg*. 2023;33(6):1790-1796. doi:10.1007/s11695-023-06603-5

80. Carmancion KM. News Verifiers Showdown: A Comparative Performance Evaluation of ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard in News Fact-Checking. *arXiv*. Preprint posted online June 18, 2023. doi:10.1109/FNWF58287.2023.10520446

81. Wang G, Gao K, Liu Q, et al. Potential and limitations of ChatGPT 3.5 and 4.0 as a source of COVID-19 information: comprehensive comparative analysis of generative and authoritative information. *J Med Internet Res*. 2023;25:e49771. doi:10.2196/49771

82. Hassan AM, Rajesh A, Asaad M, et al. Artificial intelligence and machine learning in prediction of surgical complications: current state, applications, and implications. *Am Surg*. 2023;89(1):25-30. doi:10.1177/00031348221101488

SUPPLEMENT 1.

eAppendix 1. Chat GPT Training and Writing Prompts
eAppendix 2. 10-Point and 20-Point Scale Rubrics
eAppendix 3. Chat GPT Grading Prompts

SUPPLEMENT 2.

Data Sharing Statement