Evaluation of artificial intelligence-generated layperson's summaries from abstracts of vascular surgical scientific papers

Walter Dorigo, MD,[a] Elena Giacomelli, MD, PhD,[a] Cristiano Calvagna, MD,[b] Filippo Griselli, MD,[b] Sara Speziali, MD,[a] Aaron Thomas Fargion, MD,[a] Sandro Lepidi, MD,[c] Raffaele Pulli, MD,[a] *and* Mario D'Oria, MD,[c] *Florence and Trieste, Italy*

ABSTRACT

Background: This study aimed to assess the efficacy of ChatGPT 3.5, an artificial intelligence (AI) language model, in generating readable and accurate layperson's summaries from abstracts of vascular surgery studies.

Methods: Abstracts from four leading vascular surgery journals published between October 2023 and December 2023 were used. A ChatGPT prompt for developing layperson's summaries was designed based on established methodology. Readability measures and grade-level assessments were compared between original abstracts and ChatGPT-generated summaries. Two vascular surgeons evaluated a randomized sample of ChatGPT summaries for clarity and correctness. Readability scores of original abstracts were compared with ChatGPT-generated layperson's summaries using a *t* test. Moreover, a subanalysis based on abstract topics was performed. Cohen's kappa assessed interrater reliability for accuracy and clarity.

Results: One-hundred fifty papers were included in the database. Statistically significant differences were observed in readability measures and grade-level assessments between original abstracts and AI-generated summaries, indicating improved readability in the latter (mean Global Readability Score of 36.6 ± 13.8 in the original abstract and of 50.5 ± 11.1 in the AI-generated summary; $P < .001$). This trend persisted across abstract topics and journals. Although one physician found all summaries correct, the other noted inaccuracies in 32% of cases, with mean rating scores of 4.0 and 4.7, respectively, and no interobserver agreement (k value $= -0.1$).

Conclusions: ChatGPT demonstrates usefulness in producing patient-friendly summaries from scientific abstracts in vascular surgery, although the accuracy and quality of AI-generated summaries warrant further scrutiny. (JVS-Vascular Insights 2024;2:100107.)

Keywords: Vascular Surgery; Artificial Intelligence; Large language models; Patient education; Scientific dissemination

With the growing interest of the world's population in learning about news, novelties, and technological advances in the world of medicine and health, the ability to make complex scientific research clear and accessible becomes increasingly important.[1] The European Union, for example, specifically requires that all randomized controlled trials (RCTs) be accompanied by a layperson's summary.[2] Although there are numerous pointers and strategies for making the summary simple, clear, readable, and understandable by an extremely large and diverse audience,[1,3] recent studies show that layperson's summaries may not meet the recommended reading level for medical literature.[4] The use of artificial intelligence (AI) and large language models (LLMs) has been proposed as a tool to generate summaries from scientific papers.[5–8] Open AI's ChatGPT[9] is a LLM with the ability to analyze data and provide accurate summaries of information.[10] The few data in the literature about the capability of ChatGPT to generate layperson summaries are conflicting: Eppler et al[5] demonstrated excellent results in creating layperson summaries from abstracts of leading urological scientific journals, and Kuckelman et al[6] showed that AI was effective in generating patient-friendly summaries of musculoskeletal radiology reports. In contrast, Hwang et al[7] showed that AI-generated summaries of abstracts from RCTs of various specialties, even if more readable than the originals, had significantly lower quality. Similarly, Haidar et al[8] pointed out substantial unreliability and inaccuracy of ChatGPT-generated summaries of information documents created by the UK's Vascular Society and reserved for patients with vascular disease of surgical interest.

Despite the relative abundance of papers on the topic, to the best of our knowledge, there are no studies in the

From the Division of Vascular Surgery, Department of Experimental and Clinical Medicine, University of Florence, Florencea; Division of Vascular and Endovascular Surgery, Cardio-Thoracic-Vascular Department, Cattinara Hospital ASUGI, Trieste[b]; and the Division of Vascular and Endovascular Surgery, Department of Clinical Surgical and Health Sciences, University of Trieste.[c]

Correspondence: Walter Dorigo, MD, Division of Vascular Surgery, Department of Experimental and Clinical Medicine, University of Florence, Largo Brambilla 3, 50134 Florence, Italy (e-mail: walter.dorigo@unifi.it).

The editors and reviewers of this article have no relevant financial relationships to disclose per the Journal policy that requires reviewers to decline review of any manuscript for which they may have a conflict of interest.

2949-9127

https://doi.org/10.1016/j.jvsvi.2024.100107

Table I. ChatGPT prompt used to write layperson's summaries from vascular scientific abstracts

Translate the preceding abstract into a layperson summary that is understandable by or below a sixth-grade level, incorporating the following elements if available
Population of subjects/participants
Aim of the study
Results of the study
Comments on outcome(s) of the study
Conclusion supported by the findings
Indication if follow-up studies are foreseen
Moreover, the summary should adhere to the subsequent guidelines
Mean sentence length less than 20 words
Proportion of passive verbs <10%
Spell out any acronym
For any mention of medication, treatment, health-related outcome, or anything else medically related that the general public or patient might not understand, please explain it, put it in context, and/or define it.

literature concerning the use of AI in creating layperson's summaries of the results of scientific works in the vascular surgical field.

The aim of the present study was to evaluate the capability of ChatGPT in giving providers a tool to generate summaries derived from vascular surgery studies for those most pertinent to each patient. Moreover, we compared ChatGPT's proficiency in generating patient summaries with those crafted by academic authors.

METHODS

Article selection. We collected the articles published in four leading vascular surgery journals from October 2023 to December 2023. The selected journals were the *European Journal of Vascular and Endovascular Surgery* (Elsevier NV), the *Journal of Vascular Surgery* (Elsevier NV), the *Journal of Vascular Surgery Venous and Lymphatics Disorders* (Elsevier NV), and the *Journal of Endovascular Therapy* (Sage Publishing). For consistency assurance, the abstracts obtained were cross-referenced with those documented in PubMed. Editorials, commentaries, letters to editor, case reports and short presentations, special communications, image articles and historical articles were excluded from the selection. The abstracts of the selected papers were then exported into a Microsoft Excel database.

Prompt generation and layperson's summary creation using ChatGPT. To convert an original scientific abstract input into a ChatGPT-generated layperson's summary, we followed the method reported by Eppler et al.[5] Briefly, we created a prompt that enhances readability for the general public while adhering to accuracy and clarity standards outlined in the "Good Lay Summary Practice" guidelines, endorsed by the Clinical Trials Expert Group under the European Commission in 2021.[3] Following recommendations from the Irish National Adult Literacy Agency and Plain English UK guidelines, the prompt was designed to use language that is easily comprehensible. This includes maintaining a mean sentence length of <20 words and ensuring that passive verbs constitute <10% of the content.[4] Moreover, we refined the prompt trying to obtain the lowest possible variability, considering that ChatGPT outputs are stochastic for inherent nature of GPTs (generative pretrained transformers).[9] The prompt is presented in Table I. Also the method for producing each ChatGPT (version 3.5)-generated patient summary output was borrowed from the above article.[5] We established a fresh ChatGPT window and transferred the original into the ChatGPT window. Then we appended the "Layperson prompt" after the abstract and pressed enter. We recorded the duration it took for ChatGPT to generate the output, from the moment of pressing enter until completion and finally we transferred the layperson's summary output generated by ChatGPT into the database. At the very beginning of our analysis, we generated three layperson's summaries for each abstract; however, this procedure was stopped when noting that such an iterative process did not improve any of the examined metrics.

Assessment of readability. All measures of readability and grade-level assessments (RR-GLIs) were computed automatically using the Web FX tool, as conducted previously.[4,5,8] This online platform provides evaluations such as the Global Readability Score (GS), Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Score, Smog Index, Coleman-Liau Index (CLI), and Automated Readability Index (ARI), which are widely recognized metrics.

The Flesch Reading Ease (which has equal values to the GS) is based on a ranking scale from 0 to 100, with higher values indicating higher readability, while the Flesch-Kincaid Grade Level shows the required US education to be able to understand a text. The Gunning Fog Score estimates the years of formal education needed to comprehend text on the first reading, and the Smog Index estimates the years of education a person needs to comprehend writing. Finally, the CLI shows the US school level a person needs to be to understand the text, and the ARI is designed to measure how easy a text is to understand. Although a higher score indicates easier readability for GS and the Flesch-Kincaid Reading Ease, the other indices correlate a lower score with easier readability.[11]

RR-GLIs were ascertained for both the ChatGPT layperson's summary and the original abstract to facilitate comparison.

Evaluation by physicians of the summaries. Two vascular surgeons (M.D. and E.G.) belonging to two different academic hospitals, with >10 year academic experience, independently reviewed a randomized selection (15%) of ChatGPT-generated summaries for clarity and correctness. They used a 5-point Likert scale to rate the summary in comparison to the original abstract, scoring the introduction, methods, results, and conclusions based on how well they reflected the basic findings and conclusions of the original study. The score rated from 1 (completely inaccurate) to 5 (completely accurate). A mean score of ≥4 was considered as the cut-off to define the layperson's summary accurate and complete.

Statistical analysis. A comparison of readability scores was undertaken between the original abstracts and layperson summaries generated by ChatGPT, using a *t* test. Inter-rater reliability for evaluating the accuracy and clarity of the ChatGPT-generated summaries was determined using Cohen's k. The scores were presented as means with standard deviations. A subanalysis of the scores was conducted after dividing the abstracts according to the generic topic they addressed. The issues identified were as follows: cerebrovascular disease, thoracic aorta disease, abdominal aortic aneurysm, peripheral artery obstructive disease, venous and lymphatic disease, visceral artery disease, vascular access for hemodialysis, and other topics (vascular trauma, basic science research and health and university policy study). All statistical analyses were conducted using SPSS software (IBM, Armonk, NY). A significance level of <.05 (two-tailed) was chosen to establish statistical significance.

RESULTS

Article characteristics. One-hundred fifty papers were included in the database, 69 (46%) from the *Journal of Vascular Surgery*, 36 (24%) from the *European Journal of Vascular and Endovascular Surgery*, 26 (17%) form the *Journal of Endovascular Therapy*, and 19 (13%) from the *Journal of Vascular Surgery Venous and Lymphatics Disorders*. Sixteen articles were reviews with or without a meta-analysis, and two articles were subgroup analyses from previously published RCTs. We did not find any article with a patient summary. In Table II, we report the topics of the articles on the basis of the classification as specified.

Table II. Topics and characteristics of the included abstracts

Topic	No. (%)	Review	RCT
Cerebrovascular disease	9 (6)		
Thoracic aorta disease	23 (15)	4	
Abdominal aortic aneurysm	38 (25)	2	
Peripheral artery obstructive disease	38 (25)	3	1
Venous and lymphatic disease	26 (17.5)	5	1
Visceral artery disease	4 (3)	1	
Vascular access for hemodialysis	5 (3.5)		
Others	7 (5)	1	

RCT, Randomized controlled trial.

Readability and grade-level assessment results of the ChatGPT-generated layperson's summaries. The mean time for ChatGPT to generate the layperson's summary was 7.5 ± 1.3 seconds. The cumulative analysis of all 150 abstracts showed a statistically significant difference for the majority of RR-GLIs between the ChatGPT-generated layperson's summaries and the original abstracts (Table III), with a significant improvement of readability and grade-level parameters among ChatGPT-generated summaries. Only the CLI score was similar between the two groups; the ARI score was higher in AI-generated layperson summaries. The results remained the same when analyzing the readability scores on the basis of the topic of the abstract, except for papers dealing with cerebrovascular disease, where, despite a mild improvement of the GS for the ChatGPT-generated summaries, we found a slight impairment of all other indicators of readability and grade-level parameters in comparison with the original abstract (Table IV). In Table V, we report the results of the RR-GLIs on the basis of the selected journal, without any significant differences with respect to the outcomes in the whole study group.

Quality assessment of ChatGPT-generated layperson's summaries. There were significant interobserver differences in the evaluation of the correctness and clarity of the selected generated output. The layperson's summaries were rated correct in 100% of the cases by a physician, whereas the other physician found them not correct (score <4) in 32% of the cases. The mean rates for the whole summary were 4.0 ± 0.5 and 4.7 (standard error, 0.3), without any interobserver agreement (k value = −0.1). When separately analyzing the rating of the sections of the layperson's summaries, the mean values provided by the two reviewers were 4.1 and 4.4 (*P* = .09; k = 22%) for the introduction, 3.8 and 4.7 (*P* = .04; k = −0.1) for the methods, 3.8 and 4.9 (*P* = .2; k = −0.02) for the results, and 4.1 and 4.9 (*P* = .3; k = −0.08) for the conclusions.

DISCUSSION

Readability of vascular surgical scientific literature. The results of the present study show that the readability of the scientific literature in the field of vascular surgery is generally low, even in those parts, such as the abstract, that are the gateway to the scientific paper being too often the only part of the research to be read by the nonspecialist (or nonphysician) audience and to be easily accessible during web searches. In our study, the mean value of the GS was 36.6, which is considered a value applying to texts difficult to read. This is true not only

Table III. Readability scores of both original abstracts and ChatGPT-generated layperson's summaries

Readability index	Original abstract, mean ± SD	ChatGPT summary, mean ± SD	*P* value
GS	36.6 ± 13.8	50.5 ± 11.1	<.001
Flesch reading ease	36.6 ± 13.8	50.5 ± 11.1	<.001
Flesch-Kincaid grade level	11.7 ± 2.8	10.9 ± 1.9	.01
Gunning Fog score	14.5 ± 2.9	13.7 ± 2.4	.01
Smog index	10.6 ± 2.1	9.8 ± 1.8	.003
CLI	13.7 ± 2.7	13.7 ± 1.8	.8
ARI	9.3 ± 3.8	11.8 ± 2.1	<.001

ARI, Automated readability index; *CLI*, Coleman-Liau index; *GS*, Global readability score; *SD*, standard deviation.

for the vascular literature; similar results were reported in urology.[5] In the present study, we also analyzed the readability of the abstracts referring to different vascular topics, and we found that studies dealing with abdominal aortic aneurysm and those discussing vascular trauma, basic science research, and health and university policy had the lowest readability (GS of approximately 25, indicating a text very difficult to read and best understood by university graduates).

Role of the AI in generating readable layperson's summaries. A significant point highlighted by our study is that the use of LLMs proved useful in improving the readability of scientific research results in the field of vascular surgery, as summarized in the abstracts of published papers. Indeed, the ChatGPT-generated layperson's summaries showed an improvement in all the indicators of readability and grade level parameters in comparison with the original abstracts. The mean value of GS in the layperson's summaries was >50, which represents the threshold to define a text fairly difficult to read and it is in line with previously reported studies.[5] However, these values are still far from those that identify texts that are easy to read and understand (usually defined as GS values of >60), and this in some ways still represents a limitation with which AI must contend. Furthermore, by analyzing the performance of ChatGPT according to the topic of the original abstract, it was possible to identify some topics in which layperson's summaries, although improved in absolute readability, still presented GS values of <50, thus remaining difficult to read in a general sense. The values of ARI were even worse than those of the original abstract: this metric assesses the US grade level required to read a piece of text, but, in contrast with the other formulas, rather than counting syllables, it counts characters; moreover, it also counts sentences, and this sets it apart from some other formulas, to the point that in some studies this parameter is not considered among those used to define the readability of a text.[7]

Quality of the AI-generated layperson's summaries. A significant finding in the present study is, in our opinion, the very low value of interobserver agreement recorded when assessing the scientific correctness and clarity of a random sample of ChatGPT-generated layperson's summaries. The two reviewers were both academic physicians, with a large experience in scientific research and in clinical activities, as well; yet, one of them considered the layperson's summaries to be correct, clear, and well-matched to the original abstract, while the other considered one-third of them to be insufficient, thus deeming the quality of the information it provided to be unacceptable. Specifically, the main differences between the two reviewers' assessments were in the methods, results, and conclusions sections; some degree of agreement can be detected in the introduction section. This result, although, in our opinion relevant, must still be evaluated in the light of the judging methods used. For simplicity, we used a Likert scale, which probably does not guarantee sufficient objectivity in assessing the accuracy of AI-generated texts. Hwang et al[7] created an overall quality score for analyzing AI-generated summaries from RCTs, based on the adherence to the 18-item CONSORT-A checklist.[12] However, because this score was constructed specifically to evaluate RCTs, it was not consistently applicable to the abstracts analyzed in our study.

In any case, this difference in the evaluation of the quality of the AI-generated layperson's summaries somewhat reflects the controversy in the literature, well-exemplified in the works of Eppler et al,[5] who emphasized the possibility of using ChatGPT to create comprehensible and precise summaries of scientific abstracts for patients, and Hwang et al,[7] who conversely showed that ChatGPT performed inferiorly to the authors in generating good-quality scientific abstracts.

Even if the difference we found among our two observers in the present study could depend on a different scientific sensitivity or a possible underlying bias related to distrust (or enthusiasm) toward the use of AI, it is, in our opinion, worth noting. If the summaries have elicited so many different impressions from two expert and skilled physicians, one wonders what different and likely contradictory messages they will be able to send to a wide audience of ordinary citizens, who are completely unaware of such complex and technical topics. In contrast, it should be noted that, despite the lack of agreement between the two reviewers, the average score given to the layperson's summaries was nevertheless ≥4 in both cases and, therefore, still sufficient.

Possible role of ChatGPT in the daily practice. On the basis of our results, it is our opinion that AI should be

Table IV. Readability scores of both original abstracts and ChatGPT layperson's summaries on the basis of the topic

Readability index per topic	Original abstract, mean ± SD	ChatGPT summary, mean ± SD	*P* value
Cerebrovascular disease			
GS	42.7 ± 11.0	47.5 ± 11.1	.05
Flesch reading ease	42.7 ± 11.0	47.5 ± 11.1	.05
Flesch-Kincaid grade level	10.5 ± 2.0	11.7 ± 2	.01
Gunning Fog score	13.2 ± 2.0	14.4 ± 2	.04
Smog index	9.7 ± 1.4	10.3 ± 1.8	.1
CLI	12.8 ± 2.5	13.5 ± 1.6	.1
ARI	8 ± 3.1	12.4 ± 1.9	<.001
Thoracic aorta disease			
GS	34 ± 11.9	52.1 ± 8.8	<.001
Flesch Reading ease	34 ± 11.9	52.1 ± 8.8	<.001
Flesch-Kincaid grade level	12.2 ± 2.6	10.8 ± 2.5	.04
Gunning Fog score	14.9 ± 3.0	13.5 ± 1.8	.08
Smog Index	11 ± 2.1	9.6 ± 1.4	.01
CLI	14.6 ± 2.3	13.3 ± 1.7	.02
ARI	10.4 ± 3.6	11.7 ± 2	.07
Abdominal aortic aneurysm			
GS	24.8 ± 6.8	46.5 ± 9.5	.008
Flesch reading ease	24.8 ± 6.8	46.5 ± 9.5	.008
Flesch-Kincaid grade level	13.6 ± 2.0	11.7 ± 1.9	.1
Gunning Fog score	16.4 ± 2.5	15 ± 1.5	.1
Smog Index	11.8 ± 1.6	11 ± 1.3	.2
CLI	16.6 ± 1.4	14 ± 1.6	.03
ARI	12.1 ± 3.1	12.6 ± 3.1	.7
Peripheral artery obstructive disease			
GS	33.5 ± 15.0	49.2 ± 10.1	<.001
Flesch reading ease	33.5 ± 15.0	49.2 ± 10.1	<.001
Flesch-Kincaid grade level	12.2 ± 3.1	11.1 ± 1.6	.04
Gunning Fog score	15.1 ± 3.0	13.9 ± 1.9	.05
Smog Index	10.9 ± 2.4	10.1 ± 2.7	.08
CLI	13.9 ± 2.8	14 ± 1.8	.7
ARI	9.7 ± 4.2	12.1 ± 1.9	.004
Venous and lymphatic disease			
GS	39.4 ± 13.5	56.1 ± 11.2	<.001
Flesch Reading Ease	39.4 ± 13.5	56.1 ± 11.2	<.001
Flesch-Kincaid grade level	11.4 ± 2.8	9.8 ± 2	.04
Gunning Fog score	14.4 ± 2.7	12.4 ± 2.6	.02
Smog Index	10.5 ± 2.0	8.9 ± 1.9	.01
CLI	13.1 ± 2.6	13.4 ± 2.0	.6
ARI	9.0 ± 4.0	10.8 ± 2.3	.06
Visceral artery disease			
GS	31.8 ± 21.3	39 ± 21	.3
Flesch reading ease	31.8 ± 21.3	39 ± 21	.3
Flesch-Kincaid grade level	13.3 ± 5.2	12.7 ± 4	.4
Gunning Fog score	15.9 ± 5.1	15.7 ± 5.3	.8
Smog Index	11.6 ± 3.6	11.3 ± 3.8	.9
CLI	13.9 ± 3.2	14.4 ± 2.4	.8
ARI	11.3 ± 7.1	12.8 ± 4.2	.7
Vascular access for hemodialysis			
GS	37.1 ± 10.8	54.2 ± 9.9	.01
Flesch reading ease	37.1 ± 10.8	54.2 ± 9.9	.01
Flesch-Kincaid grade level	11.4 ± 1.7	10 ± 1.5	.08
Gunning Fog score	14.4 ± 2.2	12.6 ± 1.6	.05
Smog Index	10.4 ± 1.1	9.2 ± 1.3	.09
CLI	12.8 ± 2.0	13.4 ± 1.6	.2
ARI	8.2 ± 2.1	10.6 ± 1.7	.08
Others			
GS	24.8 ± 6.8	46.5 ± 9.5	.008
Flesch reading ease	24.8 ± 6.8	46.5 ± 9.5	.008
Flesch-Kincaid grade level	13.6 ± 2.0	11.7 ± 1.9	.06
Gunning Fog score	16.4 ± 2.5	15 ± 1.5	.1
Smog Index	11.8 ± 1.6	11 ± 1.3	.2
CLI	16.6 ± 1.4	14 ± 1.6	.03
ARI	12.1 ± 3.1	12.6 ± 2.1	.7

ARI, Automated readability index; *CLI*, Coleman-Liau index; *GS*, Global readability score; *SD*, standard deviation.

(Continued)

Table IV. Continued.

Table V. Readability scores of both original abstracts and ChatGPT layperson's summaries on the basis of the selected journals

Readability index per topic	Original abstract, mean ± SD	ChatGPT summary, mean ± SD	*P* value
European Journal of Vascular and Endovascular Surgery			
GS	41.7 ± 12.5	48.6 ± 10.6	.01
Flesch reading ease	41.7 ± 12.5	48.6 ± 10.6	.01
Flesch-Kincaid grade level	10.5 ± 2.2	11.6 ± 2.1	.01
Gunning Fog Score	13.3 ± 2.3	14.6 ± 2.5	.01
Smog Index	9.7 ± 1.5	10.4 ± 1.7	.04
CLI	12.8 ± 2.5	13.6 ± 1.5	.07
ARI	7.7 ± 2.9	12.5 ± 2.2	<.001
Journal of Vascular Surgery			
GS	34 ± 14.5	49.3 ± 10.6	<.001
Flesch reading ease	34 ± 14.5	49.3 ± 10.6	<.001
Flesch-Kincaid grade level	12.1 ± 3	11.1 ± 1.8	.002
Gunning Fog Score	14.8 ± 3	13.8 ± 2.2	.004
Smog Index	10.9 ± 2.2	10 ± 1.8	.005
CLI	14 ± 2.9	13.9 ± 1.8	.4
ARI	9.8 ± 4.1	11.9 ± 1.9	<.001
Journal of Endovascular Therapy			
GS	36.3 ± 13.8	49.1 ± 10.2	<.001
Flesch reading ease	36.3 ± 13.8	49.1 ± 10.2	<.001
Flesch-Kincaid grade level	11.8 ± 1.1	11.1 ± 1.6	.2
Gunning Fog Score	15 ± 2.9	14 ± 1.7	.1
Smog Index	10.7 ± 2.1	10 ± 1.4	.1
CLI	14.3 ± 2.5	13.7 ± 1.8	.1
ARI	10 ± 3.6	12 ± 1.9	.03
Journal of Vascular Surgery Venous and Lymphatics Disorders			
GS	36.7 ± 11.5	60.1 ± 10.1	<.001
Flesch reading ease	36.7 ± 11.5	60.1 ± 10.1	<.001
Flesch-Kincaid grade level	12 ± 2.5	9 ± 1.7	<.001
Gunning Fog Score	14.9 ± 2.6	11.3 ± 1.9	<.001
Smog Index	10.9 ± 2.1	8.1 ± 1.5	<.001
CLI	13.4 ± 2.3	12.8 ± 2.1	.2
ARI	9.8 ± 3.7	10 ± 2.2	.4

ARI, Automated readability index; *CLI*, Coleman-Liau index; *GS*, Global readability score; *SD*, standard deviation.

used cautiously in disseminating the results of complex scientific research, which is not easily interpretable and may involve advanced technologies not commonly understood by the general population. It is plausible that the use of such tools could be very useful in making simple, generic content accessible, which generally regards the promotion and preservation of health in the population. In fact, recently Mondal et al[13] evaluated the effectiveness of ChatGPT in providing answers to patients' queries about lifestyle-related diseases or disorders and found that the responses were accurate and provided adequate guidance for patient management. Coraci et al[14] emphasized the ability of the LLMs to develop patient-directed evaluation tools and disease-specific questionnaires. In contrast, Haidar et al,[8] in the only published article dealing with the role of AI in disseminating knowledge about vascular surgery, showed that AI-generated information about vascular surgical procedures was poor in both the readability of text and the quality of information, and confirmed the primary role of vascular physicians in developing the empowerment and the engagement of patients with diseases requiring vascular interventions. Similar outcomes were reported by Melissano et al[15] with reference to aortic disease. Moreover, Dhar et al[16] found that ChatGPT-generated postoperative instructions for patients after tonsillectomy provided eloquently written inaccurate information, leading to patients using AI-generated medical advice contrary to physician advice. Therefore, careful monitoring by physicians and academics is absolutely necessary before disseminating AI-generated specific medical guidance.

For such reasons, an attitude characterized by substantial interest in this new technology, combined, however, with a high degree of caution and scientific rigor, is in our view even more necessary at a time in history when, in the literature, papers are beginning to appear comparing the role of humans vs AI in the writing of cover letters and more or less extensive portions of clinical notes and research papers.[17,18] The large-scale diffusion of AI in the health care world will inevitably tend to shift the ethical dilemma[19] of its use from how to preserve academic integrity to the need to protect the citizen and patient from receiving inaccurate, often misleading, and sometimes dangerous information. It is on this ground that future studies must move, to understand how physicians would be most likely to use ChatGPT in practice, and how it functions in both generating patient instructions for procedures and medications and making simple and understandable the results of scientific advancements.

Limitations and strengths of the study. The main limitation of this study is the use of the Version 3.5 of ChatGPT; nowadays GPT-4 is available online at a monthly cost ranging between $US20 and $US 25, and it has been reported to outperform the previous version "to create increasingly sophisticated and capable language models,"[10] even in accurately responding to complex vascular surgery questions.[20] We opted for the free-to-access version, because it may be available more easily to patients and hence more likely to reflect the experiences of people seeking information on vascular surgery scientific studies through LLMs. Moreover, considering that responses within the same version of ChatGPT might vary slightly when queried on different occasions, our scores could differ if queries were made at different times. Another limit was that the quality assessment of the AI-generated summaries was performed using a Likert scale without directly relying on consolidated recommended standards,[7] as previously mentioned. Finally, we used the English version of the ChatGPT, all our generated summaries were in English, and the RR-GLIs referred to the Anglo-Saxon (namely, American) literature. This factor makes our study poorly applicable to our daily practice in Italian hospitals. It is our aim to enlarge the study by translating in Italian both the original abstracts and the AI-generated layperson summaries to test their readability in a large sample of Italian citizens referring to our institutions. At this time, it is our intention, in case of significant differences among experts in assessing the quality of AI-generated texts, to find a third qualified reviewer and see who he/she aligns with. The main strength of the study is that it is the first to analyze the effectiveness of AI in generating abstracts of scientific papers from the field of vascular surgery; moreover, we performed a deep analysis of the performances of the ChatGPT-generated summaries on the basis of the different treated topics, and this is again a novelty in the scenario of papers dealing with the role of AI in medical sciences.

CONCLUSIONS

This study demonstrates that ChatGPT is a useful tool for creating readable, patient-friendly layperson summaries from abstracts of scientific papers in the field of vascular surgery, improving their readability and RR-GLIs. However, the readability of AI-generated summaries remained intermediate and for no topic was the AI able to generate summaries easy to read and to understand on the basis of the RR-GLIs. Moreover, the level of accuracy, clarity, and quality of the AI-generated layperson summaries was controversial, in our opinion making the intermediary role exercised by the physician between computer outcome and patient still fundamental. Further studies using more advanced versions of LLMs and standardized criteria for quality assessment are necessary.

The authors are willing to make their data, analytic methods, and study materials available to other researchers. All data of the study are available from the corresponding author. The research was not preregistered. The authors certify that there is no conflict of interest with any financial organization regarding the material discussed in the manuscript and that the study is unfunded.

AUTHOR CONTRIBUTIONS

Conception and design: WD, SS, RP, MD
Analysis and interpretation: WD, EG, FG, AF, SL, MD
Data collection: WD, EG, CC
Writing the article: WD, SS
Critical revision of the article: WD, EG, CC, FG, AF, SL, RP, MD
Final approval of the article: WD, EG, CC, FG, SS, AF, SL, RP, MD
Statistical analysis: WD
Obtained funding: Not applicable
Overall responsibility: WD

DISCLOSURES

None.

REFERENCES

1. Maurer M, Siegel JE, Firminger KB, Lowers J, Dutta T, Chang JS. Lessons learned from developing plain language summaries of research studies. *Health Lit Res Pract*. 2021;5:e155–e161.
2. European Union. Regulation (EU) No 536/2014 of the European parliament and of the council of 16 april 2014 on clinical trials on medicinal products for human use, and repealing directive 2001/20/EC text with EEA relevance. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri5celex%3A32014R0536.
3. Good Lay Summary Practice. 2023. https://health.ec.europa.eu/system/files/2021-10/glsp_en_0.pdf.
4. Shiely F, Daly A. Trial lay summaries were not fit for purpose. *J Clin Epidemiol*. 2023;156:105–112.
5. Eppler MB, Ganjavi C, Knudsen JE, et al. Bridging the gap between urological research and patient understanding: the role of large language models in automated generation of layperson's summaries. *Urol Pract*. 2023;10:436–443.
6. Kuckelman IJ, Wetley K, Yi PH, Ross AB. Translating musculoskeletal radiology reports into patient-friendly summaries using ChatGPT-4. *Skeletal Radiol*. 2024;53:1621–1624.
7. Hwang T, Aggarwal N, Khan PZ, et al. Can ChatGPT assist authors with abstract writing in medical journals? Evaluating the quality of scientific abstracts generated by ChatGPT and original abstracts. *PLoS One*. 2024;19:e0297701.
8. Haidar O, Jaques A, McCaughran PW, Metcalfe MJ. AI-generated information for vascular patients: assessing the standard of procedure-specific information provided by the ChatGPT AI-language model. *Cureus*. 2023;15:e49764.
9. https://openai.com/blog/chatgpt.
10. Bhattacharya K, Bhattacharya AS, Bhattacharya N, Yagnik VD, Garg P, Kumar S. ChatGPT in surgical practice—a new kid on the block. *Indian J Surg*. 2023;85:1346–1349.
11. https://www.webfx.com/tools/read-able/.
12. Hopewell S, Clarke M, Moher D, et al, CONSORT Group. CONSORT for reporting randomized controlled trials in journal and conference abstracts: explanation and elaboration. *PLoS Med*. 2008;5:e20.
13. Mondal H, Dash I, Mondal S, Behera JK. ChatGPT in answering queries related to lifestyle-related diseases and disorders. *Cureus*. 2023;15:e48296.
14. Coraci D, Maccarone MC, Regazzo G, Accordi G, Papathanasiou JV, Masiero S. ChatGPT in the development of medical questionnaires. The example of the low back pain. *Eur J Transl Myol*. 2023;33:12114.
15. Melissano G, Tinelli G, Soderlund T. Current Artificial Intelligence Based Chatbots May Produce Inaccurate and Potentially Harmful Information for Patients With Aortic Disease. *Eur J Vasc Endovasc Surg*. 2023;67:683–684.
16. Dhar S, Kothari D, Vasquez M, et al. The utility and accuracy of ChatGPT in providing post-operative instructions following tonsillectomy: a pilot study. *Int J Pediatr Otorhinolaryngol*. 2024;179:111901.
17. Deveci CD, Baker JJ, Sikander B, Rosenberg J. A comparison of cover letters written by ChatGPT-4 or humans. *Dan Med J*. 2023;70:A06230412.

18. Sikander B, Baker JJ, Deveci CD, Lund L, Rosenberg J. ChatGPT-4 and human researchers are equal in writing scientific introduction sections: a blinded, randomized, non-inferiority controlled study. *Cureus*. 2023;15:e49019.

19. Miao J, Thongprayoon C, Suppadungsuk S, Garcia Valencia OA, Qureshi F, Cheungpasitporn W. Ethical dilemmas in using AI for academic writing and an example framework for peer review in nephrology academia: a narrative review. *Clin Pract*. 2023;14:89—105.

20. Javidan AP, Feridooni T, Gordon L, Crawford SA. Evaluating the progression of artificial intelligence and large language models in medicine through comparative analysis of ChatGPT-3.5 and ChatGPT-4 in generating vascular surgery recommendations. *JVS-Vasc Insights*. 2024;2:100049.

Submitted May 9, 2024; accepted May 29, 2024.