Title

Validation of ChatGPT 3.5 as a Tool to Optimize Readability of Patient-facing Craniofacial Education Materials.

Permalink

https://escholarship.org/uc/item/4nh1955q

Journal

Plastic and Reconstructive Surgery, Global Open, 12(2)

ISSN

2169-7574

Authors

Vallurupalli, Medha
Shah, Nikhil
Vyas, Raj

Publication Date

2024-02-01

DOI

10.1097/GOX.0000000000005575

OPEN

PRS GLOBAL OPEN

ORIGINAL ARTICLE

Business

Validation of ChatGPT 3.5 as a Tool to Optimize Readability of Patient-facing Craniofacial Education Materials

Medha Vallurupalli, BA*†
Nikhil D. Shah, MD†
Raj M. Vyas, MD, FACS†‡

Background: To address patient health literacy, the American Medical Association recommends that readability of patient education materials should not exceed a sixth grade reading level; the National Institutes of Health recommend no greater than an eigth-grade reading level. However, patient-facing materials in plastic surgery often remain at an above-recommended average reading level. The purpose of this study was to evaluate ChatGPT 3.5 as a tool for optimizing patient-facing craniofacial education materials.
Methods: Eighteen patient-facing craniofacial education materials were evaluated for readability by a traditional calculator and ChatGPT 3.5. The resulting scores were compared. The original excerpts were then inputted to ChatGPT 3.5 and simplified by the artificial intelligence tool. The simplified excerpts were scored by the calculators.
Results: The difference in scores for the original excerpts between the online calculator and ChatGPT 3.5 were not significant ($P = 0.441$). Additionally, the simplified excerpts' scores were significantly lower than the originals ($P < 0.001$), and the mean of the simplified excerpts was 7.78, less than the maximum recommended 8.
Conclusions: The use of ChatGPT 3.5 for simplification and readability analysis of patient-facing craniofacial materials is efficient and may help facilitate the conveyance of important health information. ChatGPT 3.5 rendered readability scores comparable to traditional readability calculators, in addition to excerpt-specific feedback. It was also able to simplify materials to the recommended grade levels. With human oversight, we validate this tool for readability analysis and simplification. *(Plast Reconstr Surg Glob Open 2024; 12:e5575; doi: 10.1097/GOX.0000000000005575; Published online 2 February 2024.)*

INTRODUCTION

In the new Healthy People 2030 definitions, the US Department of Health and Human Services defines personal health literacy as "the degree to which individuals have the ability to find, understand, and use information and services to inform health-related decisions and actions for themselves and others." This new definition is emblematic of the ongoing shift toward patient-centered care and shared decision-making, in which patients' preferences, needs, and values play a key role in treatment planning and informed decision-making.[1] However, the National Assessment of Adult Literacy Survey reported that 36% of US adults had basic or below-basic health literacy, and approximately 80 million adults were estimated to have limited or low health literacy.[2,3] Limited health literacy plays a detrimental role in patient understanding of health information and hinders patients from being actively involved in their care. In addition, due to the stigma of low health literacy, patients may not admit difficulties or seek assistance when needed.[4,5]

Although patient education materials, such as brochures and web pages, aim to provide patients with

Disclosure statements are at the end of this article, following the correspondence information.

Related Digital Media are available in the full-text version of the article on www.PRSGlobalOpen.com.

*From the *Keck School of Medicine of USC, Los Angeles, Calif.; †Department of Plastic Surgery, University of California, Irvine, Calif.; and ‡CHOC Children's Hospital of Orange County, Orange, Calif.*

Received for publication October 2, 2023; accepted December 11, 2023.

DOI: 10.1097/GOX.0000000000005575

relevant information, they often require an above-average reading level to understand.[6,7] A 2020 Gallup analysis found that about 54% of Americans between the ages of 16 and 74 read below a sixth grade level. To address this, the American Medical Association recommends that readability of patient education materials should not exceed a sixth grade reading level. The National Institutes of Health recommend no greater than an eigth-grade reading level.[8] However, patient-facing materials in plastic surgery often remain at an above-recommended average reading level.[9–20]

Artificial intelligence (AI) has had emerging areas of application in healthcare and has primarily been used in diagnostics.[21,22] However, recently developed AI models may have further areas of application. ChatGPT 3.5 (OpenAI, San Francisco, Calif.), a language learning model, is one of several AI interfaces that uses natural language processing to provide users with answers to queries. Craniofacial procedures often require complex reconstruction, which can be difficult to explain. Although prior studies have evaluated the readability of various patient-facing materials, to our knowledge, there are none that have attempted to use AI to analyze readability and to simplify these materials to the recommended sixth- to eigth-grade level. The purpose of this study was to evaluate ChatGPT 3.5 as a tool for optimizing patient-facing craniofacial education materials.

Takeaways

Question: Can ChatGPT 3.5 be used to analyze and simplify patient-facing craniofacial surgery materials?

Findings: Craniofacial education materials were evaluated for readability by a traditional calculator and ChatGPT 3.5. ChatGPT 3.5 rendered readability scores comparable to traditional readability calculators, in addition to providing excerpt-specific feedback. It was also able to simplify materials to the recommended grade levels established by the American Medical Association and National Institutes of Health.

Meaning: ChatGPT can be used to analyze and simplify patient education materials in craniofacial surgery, promoting efforts to increase patient health literacy.

METHODS

A total of 18 publicly available excerpts from patient-facing education materials at US academic institutions and the American Cleft Palate Craniofacial Association were selected for readability analysis through web searches of common craniofacial procedures. These excerpts included patient information regarding cleft lip, cleft palate, craniosynostosis, orthognathic surgery, and nerve decompression surgery for migraines obtained from prominent patient-facing websites. As this was a novel study for which there were no previous available data, it was not possible to estimate an effect size and perform a power analysis. Therefore, the decision to include these 18 excerpts was based on consensus between authors regarding procedures of interest and search engine prominence. All selected excerpts were from academic medical institutions and national craniofacial organizations (Table 1). A traditional online readability calculator, Readability Scoring System v1.0, was then used to calculate grade level and readability for each of the excerpts, using averaged metrics of seven readability indexes: Flesch reading ease score, Gunning fog index, Flesch-Kincaid grade level, the Coleman-Liau index, the Simple Measure of Gobbledygook index, automated readability index, and the Linsear Write formula. Each index uses varied metrics to calculate grade level scores. Examples of metrics include words per sentence, letters per word, and percentage of complex words. This score was used as the baseline score for each excerpt. Visual accompaniments could not be assessed and were removed before evaluation. Then, excerpts were analyzed by ChatGPT 3.5 for grade level and readability with the prompt, "Analyze this paragraph for readability. Provide readability analysis and include grade level." ChatGPT 3.5 was chosen because it is a free, widely available, and popular AI tool, when compared with ChatGPT 4.0, which is a paid subscription-based service and currently has an hourly message cap. To reduce barriers to access, ChatGPT 3.5 was chosen as the AI tool. Readability analysis using ChatGPT 3.5 included words per sentence, characters per word, syllables per word, and grade level. It also included a descriptive evaluation of the excerpts and written suggestions to simplify the given information. Statistical comparisons between the ChatGPT 3.5 scores and baseline scores were performed using a Wilcoxon signed rank test.

Table 1. Websites Accessed

Organization	Website	No. Excerpts
ACPA	https://acpacares.org	4
Children's Hospital of Philadelphia	https://www.chop.edu	2
Nationwide Children's	https://www.nationwidechildrens.org	1
Johns Hopkins Medicine	https://www.hopkinsmedicine.org	3
Children's Hospital of Orange County	https://www.choc.org	1
American Society of Plastic Surgeons	https://www.plasticsurgery.org	3
Mayo Clinic	https://www.mayoclinic.org	1
Boston Children's Hospital	https://www.childrenshospital.org	1
Medical University of South Carolina	https://muschealth.org	1
University of Wisconsin - Madison	https://www.surgery.wisc.edu	1

Finally, ChatGPT 3.5 was used to simplify the chosen excerpts to the recommended sixth to eigth grade reading level. Excerpts were inputted to ChatGPT 3.5 with the prompt, "Rewrite this paragraph for an 8th grader without losing information from the original paragraph." The simplified paragraphs were then read by the authors to confirm accuracy. These excerpts were then inputted to the online readability calculator for grade level and readability analysis. A Wilcoxon signed rank test was used to compare the excerpts. Statistical analyses were conducted with IBM SPSS Statistics (version 28).

RESULTS

ChatGPT 3.5 as a Readability Calculator

Of the 18 excerpts, seven excerpts had a calculator score of 10; five excerpts had a calculator score of 11; and six excerpts had a calculator score of 12. These scores represent the reading grade level needed to comprehend the materials. The mean calculator score was 10.94. The mean ChatGPT 3.5 score was 10.79. Traditional online readability calculator and ChatGPT 3.5 scores are reported in Table 1. The difference in scores between the online calculator and ChatGPT 3.5 (Table 2) were not significant ($P = 0.441$).

ChatGPT 3.5 as a Simplification Tool

The original excerpts had an average grade level score of 10.79, which was significantly greater than the maximum national recommended grade level of 8 ($P < 0.001$; 95% CI, 2.51–3.38). Of the 18 simplified excerpts, one excerpt had a calculator score of 6; three excerpts had a calculator score of 7; 13 had a calculator score of 8; and one had a calculator score of 9 (Fig. 1). The mean grade level for the simplified excerpts with the traditional readability calculator was 7.78. The mean grade level for the simplified excerpts with ChatGPT 3.5 was 7.69. Individual calculator scores are reported in Table 1. The simplified excerpts were significantly different from the original ($P < 0.001$). The mean of the simplified excerpts was 7.78, less than the maximum recommended 8.

DISCUSSION

Because 54% of Americans between the ages of 16 and 74 read at a level equivalent to, or below, sixth grade, the American Medical Association and National Institutes of Health both urge that patient education materials should not be written at a level greater than eigth grade (ideally maintaining an average between sixth and eigth grade[8]). Despite these recommendations, prior studies have elucidated the readability of patient education materials, with many studies describing that the average reading levels of education materials do not meet the national recommended average by several grade levels.[6,7,9–20] In prior studies, it is clear that many patient education materials are written above the recommended levels; however, to our knowledge, this is the first study to offer a simplified solution using AI to manage and improve patient health education materials.

In this study, we have validated the efficacy of ChatGPT 3.5 as both a readability measurement and highlighted its application as a reliable text simplification tool for craniofacial education materials. There was no statistically significant difference between the readability scores of the traditional online calculator and ChatGPT 3.5 for the 18 excerpts, validating ChatGPT 3.5 as a reliable tool for assessing the readability of patient education materials. Furthermore, ChatGPT 3.5's ability to provide detailed, excerpt-specific feedback distinguishes it from traditional readability tools, which generally provide only numerical grade level analyses. Although online calculators can provide grade level analysis, they cannot provide specific feedback. Thus, ChatGPT 3.5 may be a more efficient tool for measuring readability.

Table 2. Calculator and ChatGPT Individual Grade Scores Per Excerpt

Excerpt Number	Procedure Type	ChatGPT Original Grade Level	Calculator Original Grade Level	Simplified Grade Level
1	Cleft lip/palate	11	11	8
2	Cleft lip/palate	12	11	8
3	Cleft lip/palate	10	10	8
4	Cleft lip/palate	12	11	9
5	Cleft lip/palate	11	10	8
6	Craniosynostosis	9	10	7
7	Craniosynostosis	11	12	8
8	Craniosynostosis	10	11	8
9	Craniosynostosis	10	10	8
10	Craniosynostosis	10	10	6
11	Orthognathic surgery	10	11	8
12	Orthognathic surgery	12	12	8
13	Orthognathic surgery	9	10	7
14	Orthognathic surgery	10	10	8
15	Migraine nerve decompression	12	12	8
16	Migraine nerve decompression	12	12	8
17	Migraine nerve decompression	12	12	8
18	Migraine nerve decompression	11.3	12	7

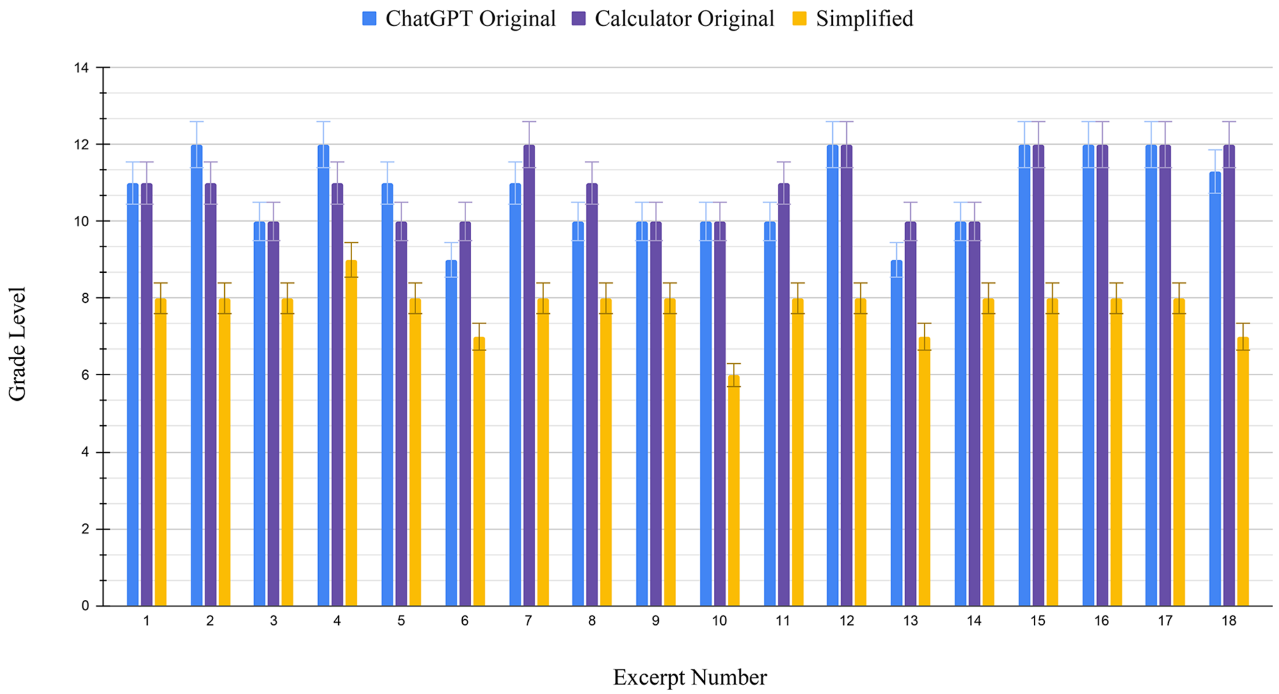

Fig. 1. Grade-level scores of original and simplified excerpts.

This feedback provided by ChatGPT 3.5 includes information about the target audience, identifies areas for potential improvement, and offers specific suggestions for simplifying complex information about craniofacial procedures. By identifying areas of potential difficulty, ChatGPT 3.5 may aid physicians in revising or curating complex information. This can streamline the process of revising patient education materials, saving clinicians time and effort while improving the overall efficacy and impact of patient education.

In addition to evaluating the readability of existing patient education materials, we found that ChatGPT 3.5 can be effectively used to simplify complex information regarding craniofacial procedures. This feature is particularly useful, given that the grade levels of the original 18 excerpts we tested significantly exceeded the national recommended eigth-grade level. Reading levels above the national average present a barrier to effective patient–provider communication, inhibit patient education and informed decision-making, and negatively impact overall health literacy. However, ChatGPT 3.5 was capable of significantly improving the readability of these materials. After simplification by the AI, the mean grade level of the revised excerpts was not only significantly lower than the originals, but also below the national eigth grade recommendation. Additionally, author review of the generated excerpts did not necessitate any changes to maintain accuracy.

This capability of ChatGPT 3.5 to simplify health-related information is critically important for revising current health materials and creating future resources that align with national readability recommendations. Lowering the reading level of patient education materials allows for better patient understanding of medical conditions and treatments, empowering patients to make informed decisions about their care. Having a clear understanding of their health allows patients to play a key role in treatment planning.[1] This analysis is important for tailoring future health materials and revising materials which currently do not meet the national recommended average (Table 3).

LIMITATIONS

Although ChatGPT 3.5 has shown significant utility, we also note significant limitations of the recently released tool. For example, when given broad statements such as, "Simplify this paragraph," ChatGPT 3.5 will provide a nonspecific response, simplifying materials broadly and often removing key information from the original input. We recommend exercising specificity in the input query to maximize the utility of the response and further human oversight to evaluate the validity, tone, and content of the simplified responses. However, also notable is ChatGPT 3.5's ability to self-correct; we also recommend providing feedback and a follow-up prompt to the chatbot if initial outputs are not up to standard. Further areas of study may evaluate the nature of the follow-up responses after unsatisfactory initial outputs and the quality and impact of the simplified information and materials. In addition, the current version of ChatGPT 3.5 has knowledge limited to information published online before 2021. Information published after 2021 is not yet accessible to the chatbot. As such, we recommend fact-checking the simplified excerpts for accuracy and supplementing with recently-published information as needed. In addition,

Table 3. Example Excerpt and Simplification

Original Excerpt	ChatGPT Simplified Excerpt
What can be done about speech problems? Speech therapy alone may be able to correct your child's speech disorder. Therapy can be extremely effective for children with mild hypernasality, an articulation disorder, or speech delay. The goal of speech therapy will be to develop good speech habits as well as to learn how to produce sounds correctly. Speech therapy alone will generally NOT correct hypernasality that is caused by moderate to severe velopharyngeal inadequacy. The type of therapy your child receives will be determined by the type of problem your child has. Furthermore, the amount of therapy your child needs will depend on the severity of the speech problem. If your child's articulation difficulties are related to a dental abnormality, the combination of articulation therapy and dental treatment can help to minimize the problem. If your cleft palate team decides that speech therapy alone will not correct your child's speech problem, there are some other options. Your child may require another palate surgery to help with speech. The two most common speech surgeries are (1) pharyngeal flap and (2) sphincter pharyngoplasty. (The surgeon may also choose to redo the original repair.) Your child's speech-language pathologist and surgeon will work together to determine the most appropriate type of surgery for your child. Talk to your surgeon about which procedure he or she intends to perform. It is important to remember that surgery is not a "quick fix." It is almost always necessary for a child to participate in speech therapy after surgery to practice correct articulation and good speech habits. Although surgery is the most frequently-chosen approach for improving velopharyngeal function, a prosthetic device may be an option for some patients. These speech aids are placed in the mouth, much like an orthodontic retainer. The two most common types are (1) the speech bulb and (2) the palatal lift. The speech bulb is designed to partially close off the space between the soft palate and the throat. The palatal lift appliance serves to lift the soft palate to a position that makes closure possible. Many professionals feel that prosthetic appliances work best in children who are at least aged 5 years. However, each patient should be evaluated on an individual basis to determine if one of these devices is appropriate for him or her.	What can be done about speech problems? Speech therapy can help a lot with fixing speech issues. It works well for kids who have mild problems, like talking through their nose too much, trouble making certain sounds, or being slow to talk. The goal of speech therapy is to teach good speaking habits and how to make sounds correctly. But sometimes, speech therapy alone might not be enough if the problem is more serious. Your child's therapy will depend on their specific problem. How much therapy they need will depend on how bad the speech problem is. Sometimes, if the problem is because of their teeth, a combination of speech therapy and dental treatment can help. If speech therapy alone doesn't fix the problem, there are other options to consider. Your child might need another surgery to help with their speech. The most common surgeries are called pharyngeal flap and sphincter pharyngoplasty. Your child's speech therapist and surgeon will work together to decide which surgery is best. It's important to know that surgery isn't a quick fix, and your child will still need speech therapy afterward to practice good speaking habits. Another option for some kids is to use a special device in their mouth, kind of like a retainer. These devices can help with the way they talk by either closing off a space in their mouth or lifting up the soft part of the roof. Not every kid needs this, so it's important to talk to the doctor to see if it's right for your child.

as a constantly developing technology, future versions of ChatGPT 3.5 require subsequent study of improvements and limitations.

As AI progresses, publicly available technology constantly evolves, and consistency may not always be guaranteed. We recommend using the tool for robust primary analysis and simplification with a subsequent human confirmation of the output to maximize time- and cost-efficiency. In short, ChatGPT 3.5 is best used as a supplement to natural intelligence.

In addition, we note limitations of our study. The presence of visual aids can play a substantial role in patient understanding of information. However, neither the online calculator nor ChatGPT 3.5 possess image or video processing abilities; therefore, supplemental content, which could affect the grade level scores, was removed. Further evaluation of these supplemental materials would more comprehensively evaluate the readability of patient materials.

CONCLUSIONS

In this study, we found that the use of ChatGPT 3.5 for simplification and readability analysis of patient-facing craniofacial materials is efficient and may help facilitate the conveyance of important health information. We note that we have only validated this tool for craniofacial procedures. Craniofacial surgery was used as a testbed for further analyses; we are currently in the process of analyzing applicability to broader areas of plastic surgery. Although further research is required regarding the impact and quality of the simplified materials, the use of AI for patient education seems promising. Craniofacial surgery encompasses many complex procedures, care instructions, and risks, which can often be misunderstood by patients. Therefore, the use of this tool to simplify information can be helpful in clinical practice, particularly for patients at risk for low health literacy.

Raj M. Vyas, MD, FACS
UC Irvine Department of Plastic Surgery
200 S. Manchester Ave, Suite 650
Orange, CA 92868
E-mail: rajv1@hs.uci.edu

DISCLOSURE

The authors have no financial interest to declare in relation to the content of this article.

REFERENCES

1. Greene SM, Tuzzio L, Cherkin D. A framework for making patient-centered care front and center. *Perm J.* 2012;16:49–53.
2. Coughlin SS, Vernon M, Hatzigeorgiou C, et al. Health literacy, social determinants of health, and disease prevention and control. *J Environ Health Sci.* 2020;6:3061.
3. Hickey KT, Masterson Creber RM, Reading M, et al. Low health literacy: implications for managing cardiac patients in practice. *Nurse Pract.* 2018;43:49–55.
4. Seo J, Goodman MS, Politi M, et al. Effect of health literacy on decision-making preferences among medically underserved patients. *Med Decis Making.* 2016;36:550–556.

5. Theiss LM, Wood T, McLeod MC, et al. The association of health literacy and postoperative complications after colorectal surgery: a cohort study. *Am J Surg*. 2022;223:1047–1052.
6. Eltorai AEM, Ghanian S, Adams CA, et al. Readability of patient education materials on the American Association for Surgery of Trauma website. *Arch Trauma Res*. 2014;3:e18161.
7. Para A, Thelmo F, Rynecki ND, et al. Evaluating the readability of online patient education materials related to orthopedic oncology. *Orthopedics*. 2021;44:38–42.
8. Rooney MK, Santiago G, Perni S, et al. Readability of patient education materials from high-impact medical journals: a 20-year analysis. *J Patient Exp*. 2021;8:2374373521998847.
9. Tran BNN, Singh M, Singhal D, et al. Readability, complexity, and suitability of online resources for mastectomy and lumpectomy. *J Surg Res*. 2017;212:214–221.
10. Seth AK, Vargas CR, Chuang DJ, et al. Readability assessment of patient information about lymphedema and its treatment. *Plast Reconstr Surg*. 2016;137:287e–295e.
11. Kiwanuka E, Mehrzad R, Prsic A, et al. Online patient resources for gender affirmation surgery: an analysis of readability. *Ann Plast Surg*. 2017;79:329–333.
12. Patel MJ, Perez BR, Zhu AQ, et al. Analysis of online patient education materials on rhinoplasty. *Facial Plast Surg Aesthet Med*. 2022;24:276–281.
13. Tiourin E, Barton N, Janis JE. Health literacy in plastic surgery: a scoping review. *Plast Reconstr Surg Glob Open*. 2022;10:e4247.
14. Vargas CR, Ricci JA, Lee M, et al. The accessibility, readability, and quality of online resources for gender affirming surgery. *J Surg Res*. 2017;217:198–206.
15. Chen AD, Ruan QZ, Bucknor A, et al. Social media: is the message reaching the plastic surgery audience? *Plast Reconstr Surg*. 2019;144:773–781.
16. Vargas CR, Kantak NA, Chuang DJ, et al. Assessment of online patient materials for breast reconstruction. *J Surg Res*. 2015;199:280–286.
17. Fanning JE, Okamoto LA, Levine EC, et al. Content and readability of online recommendations for breast implant size selection. *Plast Reconstr Surg Glob Open*. 2023;11:e4787.
18. Ricci JA, Vargas CR, Chuang DJ, et al. Readability assessment of online patient resources for breast augmentation surgery. *Plast Reconstr Surg*. 2015;135:1573–1579.
19. Barton N, Janis JE. Missing the mark: the state of health care literacy in plastic surgery. *Plast Reconstr Surg Glob Open*. 2020;8:e2856.
20. Patel AA, Joshi C, Varghese J, et al. Do websites serve our patients well? A comparative analysis of online information on cosmetic injectables. *Plast Reconstr Surg*. 2022;149:655e–668e.
21. Ebrahimian S, Digumarthy SR, Bizzo B, et al. Artificial intelligence has similar performance to subjective assessment of emphysema severity on chest CT. *Acad Radiol*. 2022;29:1189–1195.
22. Duff LM, Scarsbrook AF, Ravikumar N, et al. An automated method for artifical intelligence assisted diagnosis of active aortitis using radiomic analysis of FDG PET-CT images. *Biomolecules*. 2023;13:343.