BRIEF COMMUNICATION OPEN

Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers

Catherine A. Gao[1]✉, Frederick M. Howard[2], Nikolay S. Markov[1], Emma C. Dyer[2], Siddhi Ramesh[2], Yuan Luo[3] and Alexander T. Pearson[2]

Large language models such as ChatGPT can produce increasingly realistic text, with unknown information on the accuracy and integrity of using these models in scientific writing. We gathered fifth research abstracts from five high-impact factor medical journals and asked ChatGPT to generate research abstracts based on their titles and journals. Most generated abstracts were detected using an AI output detector, 'GPT-2 Output Detector', with % 'fake' scores (higher meaning more likely to be generated) of median [interquartile range] of 99.98% 'fake' [12.73%, 99.98%] compared with median 0.02% [IQR 0.02%, 0.09%] for the original abstracts. The AUROC of the AI output detector was 0.94. Generated abstracts scored lower than original abstracts when run through a plagiarism detector website and iThenticate (higher scores meaning more matching text found). When given a mixture of original and general abstracts, blinded human reviewers correctly identified 68% of generated abstracts as being generated by ChatGPT, but incorrectly identified 14% of original abstracts as being generated. Reviewers indicated that it was surprisingly difficult to differentiate between the two, though abstracts they suspected were generated were vaguer and more formulaic. ChatGPT writes believable scientific abstracts, though with completely generated data. Depending on publisher-specific guidelines, AI output detectors may serve as an editorial tool to help maintain scientific standards. The boundaries of ethical and acceptable use of large language models to help scientific writing are still being discussed, and different journals and conferences are adopting varying policies.

npj Digital Medicine (2023)6:75 ; https://doi.org/10.1038/s41746-023-00819-6

The release of OpenAI's free tool ChatGPT[1] on November 30, 2022 demonstrated the ability of artificial intelligence models to generate content, with articles quickly published on its possible uses and potential controversies[2–4]. Early adopters have shared their experiences on social media, with largely positive sentiments[5]. Articles are bemoaning the death of the traditional school essay assignment[4,6,7], as ChatGPT has been shown to generate high-scoring papers[8], correctly answer USMLE questions[9], and even articulate critical thinking[10]. The ethical and acceptable boundaries of ChatGPT's use in scientific writing remain unclear[11], although some publishers are beginning to lay down policies[12–14].

Large language models (LLM) are often complex neural network-based transformer models that can generate tone and content-defined text. These are trained on enormous amounts of data to predict the best next text element, which produces a product that reads naturally. ChatGPT is built on Generative Pre-trained Transformer-3 (GPT-3), which is one of the largest of these types of models, trained with 175 billion parameters[15]. These models generate coherent and fluent output, that can be difficult to distinguish from text written by humans[16,17].

Artificial intelligence (AI) has numerous applications in medical technologies[18], and the writing of biomedical research is no exception, with products such as the SciNote Manuscript Writer[19] or Writefull[20] that help with scientific writing. However, with the release of ChatGPT, this powerful LLM technology is now available to all users for free, and millions are engaging with the new technology. The user base is likely to continue to grow. Thus, there is an urgent need to determine if ChatGPT can write convincing medical research abstracts.

We gathered 50 abstracts from five high-impact journals as our control corpus of well-written abstracts. We asked ChatGPT to generate 50 scientific abstracts based on the titles and specific journals from this list (example subset in Supplementary Data 1). While all the output appeared superficially to be formatted as a scientific abstract, only 8 (16%) correctly used the headings particular to the specific journal in the prompt (e.g., *Nature Medicine's* paragraph-style without headings, as opposed to specific headings such as 'Design, Setting, and Participants' for *JAMA*, see Supplementary Note 1 for examples). The patient cohort sizes were a similar order of magnitude between the original abstracts and the generated abstracts, with a Pearson correlation of the logarithmic cohort sizes of $r = 0.76$, $p < 0.001$ (Fig. 1).

The AI output detector we used, 'GPT-2 Output Detector'[21,22], found a high probability of AI-generated output (higher % 'fake' score indicating more likely to be AI-generated text) in the generated abstracts with median [IQR] of 99.98% [12.73%, 99.98%] compared with very low probability of AI-generated output in nearly all the original abstracts with median [IQR] of 0.02% [0.02%, 0.09%] (Fig. 2a). The AI output detector had an area under the receiver operating characteristics (AUROC) curve of 0.94 for detecting generated abstracts (Fig. 2b). At the optimal cutoff maximizing sensitivity and specificity, 1.46%, the AI output detector had a sensitivity of 86% and a specificity of 94% at differentiating original versus generated abstracts.

We ran both original and generated abstracts through a free plagiarism-detection website, Plagiarism Detector[20], and a paid professional similarity checker, iThenticate[23]. For both platforms, a

[1]Division of Pulmonary and Critical Care, Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA. [2]Section of Hematology/Oncology, Department of Medicine, University of Chicago, Chicago, IL, USA. [3]Division of Health and Biomedical Informatics, Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA. ✉email: catherine.gao@northwestern.edu

higher score indicates more matching text was found. Original abstracts scored higher on the Plagiarism Detector website with median ‘plagiarized’ score 62.5% [IQR 43.25%, 84.75%] compared with generated abstracts with median ‘plagiarized’ score of 0% [IQR 0, 0] (Fig. 3a). Original abstracts also scored higher on iThenticate with median similar index of 100 [IQR 100, 100] compared with generated abstracts that had with median similarity index of 27 [IQR 19, 40.75] (Fig. 3b).

Blinded human reviewers (FMH, NSM, ECD, SR) were given a mixture of real and generated abstracts and asked to provide a binary score of real or generated (Table 1). They were able to correctly identify 68% of generated abstracts as being generated, and correctly identified 86% of original articles as being original. They incorrectly identified 32% of generated abstracts as being real, and 14% of original abstracts as being generated. Our reviewers commented that abstracts they thought were generated by ChatGPT were superficial and vague, and sometimes focused on details of original abstracts such as inclusion of Clinical Trial Registration numbers and alternative spellings of words. The AI output detector scores were not statistically different ($p = 0.45$ by MWW) between the abstracts that reviewers correctly identified as generated and ones that they failed to identify as generated (Fig. 4).

In this study, we found that both humans and AI output detectors were able to identify a portion of abstracts generated by ChatGPT, but neither were perfect discriminators. Our reviewers even misclassified a portion of real abstracts as being generated, indicating they were highly skeptical when reviewing the abstracts. The generated abstracts contained fabricated numbers but were in a similar range as the real abstracts; ChatGPT knew from its training data that studies on hypertension should include a much larger patient cohort size than studies on rarer diseases such as monkeypox.

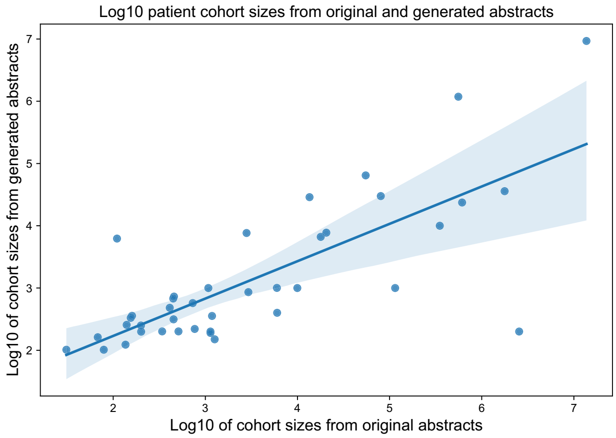

Fig. 1 Generated abstracts have a similar patient cohort size as original abstracts. Cohort sizes from original abstracts (*x*-axis) and generated abstracts (*y*-axis) plotted on a logarithmic 10 scale.

Limitations to our study include its small sample size and few reviewers. ChatGPT is also known to be sensitive to small changes in prompts; we did not exhaust different prompt options, nor did we deviate from our prescribed prompt. ChatGPT generates a different response even to the same prompt multiple times, and we only evaluated one of infinite possible outputs. We took only the first output given by ChatGPT, without additional refinement that could enhance its believability or improve its escape from detection. Thus, our study likely underestimates the ability of ChatGPT to generate scientific abstracts. The maximum input for the AI output detector we used is 510 tokens, thus some of the abstracts were not able to be fully evaluated due to their length. Our study ream reviewers knew that a subset of the abstracts they were viewing were generated by ChatGPT, but a reviewer outside this context may not be able to recognize them as written by a large language model. We only asked for a binary response from our reviewer team of original or generated and did not use a formal or more sophisticated rubric. Future studies could expand on our methodology to include other AI output detector models, other plagiarism detectors, more formalized review, as well as text from other fields outside of biomedical sciences.

We anticipate that this technology could be used in both an ethical and unethical way. Given its ability to generate abstracts with believable numbers, it could be used by organizations such as paper mills to entirely falsify research. On the other hand, the technology may be used in conjunction with a researcher’s own scientific knowledge as a tool to decrease the burden of writing and formatting. It could be used by scientists publishing in a language that is not their native language, to improve equity. However, AI models have been shown to be highly sensitive to biases in training data[24,25], and further data is needed to determine the potential for bias perpetuated by ChatGPT—especially given the overt prejudices emerging from prior language generation models[26].

We suggest clear disclosure when a manuscript is written with assistance from large language models such as ChatGPT. Though there is ongoing work to embed watermarks in AI-generated output, it is unknown when this will be standardized and robust against scrubbing efforts. Reassuringly, there are patterns that allow it to be detected by AI output detectors, although there has

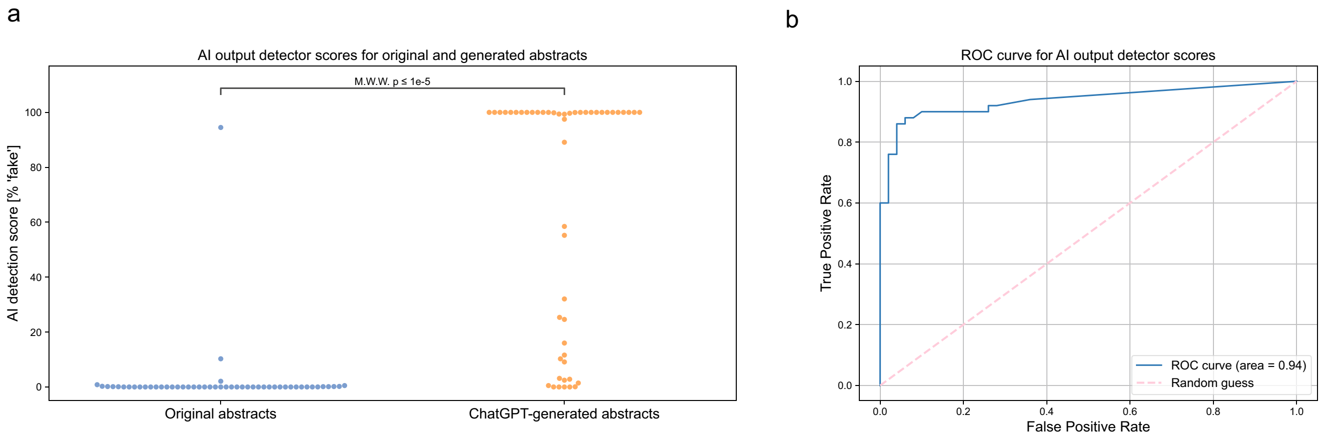

Fig. 2 Many generated abstracts can be detected using an AI output detector. a AI detection scores as [% ‘fake’] per GPT-2 Output Detector for original abstracts and generated abstracts. Higher score indicates more likely to be generated by AI. **b** The AI output detector ROC curve for discriminating between original and generated abstracts, with AUROC of 0.94.

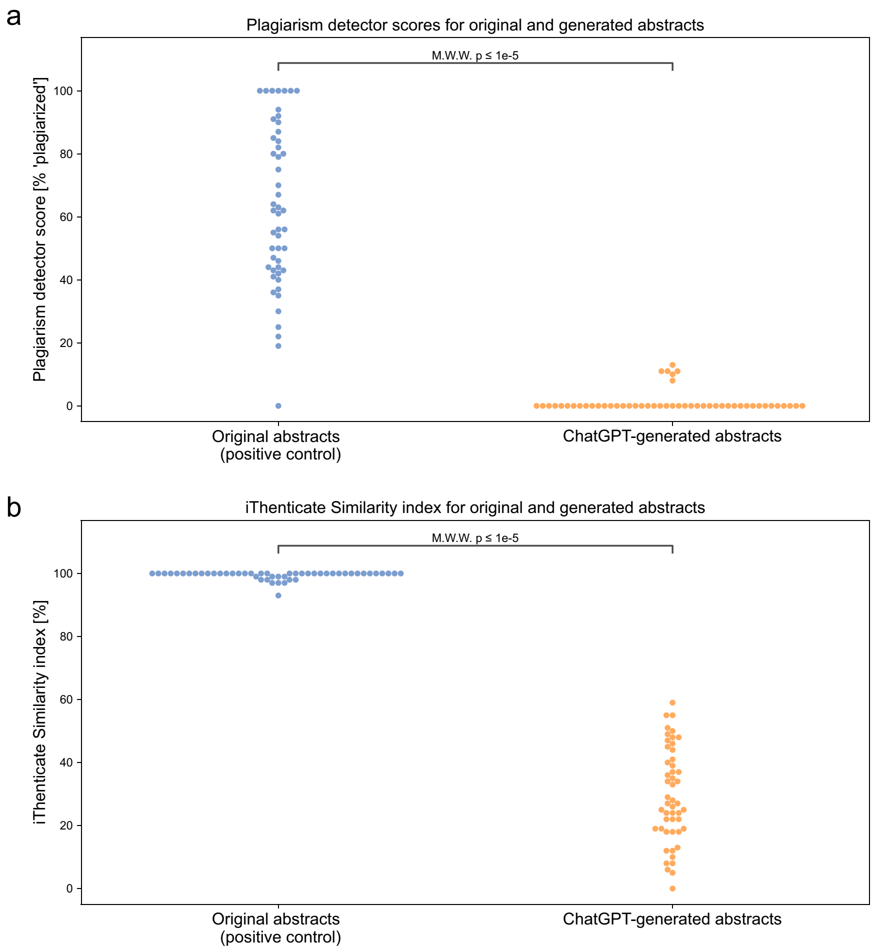

Fig. 3 Generated abstracts score lower than original abstracts on plagiarism detectors. **a** Plagiarism scores from plagiarism detector website, with higher % 'plagiarized' score indicating more matching text was found. **b** iThenticate Similarity Index for original abstracts and generated abstracts [%], with higher value meaning more similar text was found.

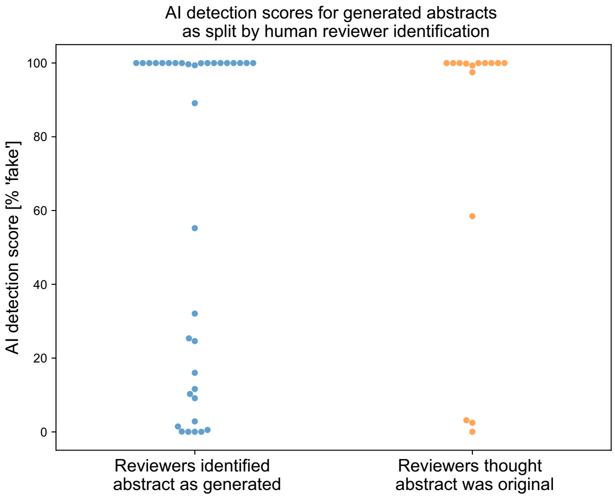

Fig. 4 Reviewers use criteria different than the AI output detector for flagging abstracts as either generated or original. The AI detection scores for generated abstracts were not significantly different ($p = 0.45$) between abstracts that human reviewers identified as generated, and those that they failed to identify as generated.

Table 1. Human reviewer scoring for whether abstracts were real or generated, along with truth.

		Truth	
		Original	Generated
Reviewer guess	Original	43	16
	Generated	7	34

been exploration of techniques to fool AI output detectors[27]. Different journals and publishers are developing their own policies on whether large language models can contribute to the writing of papers, ranging from not allowing any AI-generated text[12] to allowing its use as long as it is openly disclosed[13,14,28]. Though imperfect, AI output detectors may be one tool to include in the research editorial process, depending on the publisher or conference's guidelines.

Abstract generation by large language models such as ChatGPT is a powerful tool to create readable scientific abstracts, though includes generated data. The generated abstracts do not always alarm plagiarism-detection models, as the text is generated anew, but can often be detected using AI detection models, and sometimes identified by a person. Generated text may help alleviate the burden of writing by providing an outline for a scientist to edit but requires careful review for factual accuracy. The optimal use and ethical boundaries of AI-generated writing remain to be determined as discussion within the scientific community evolves.

METHODS

Abstract generation

We evaluated the abstracts generated by ChatGPT (Version Dec 15) for 50 scientific medical papers. We gathered titles and original abstracts from current and recent issues (published in late November and December of 2022) of five high-impact journals (*Nature Medicine, JAMA, NEJM, BMJ, Lancet*) and compared them with the original abstracts. The prompt fed to the model was 'Please write a scientific abstract for the article [title] in the style of [journal] at [link]'. Note that the link is superfluous because ChatGPT cannot browse the internet. ChatGPT's knowledge cutoff date is September 2021. We ran each prompt in a new session.

Abstract evaluation

We evaluated the ChatGPT-generated abstracts for plagiarism detection using a free web-crawling plagiarism-detection tool 'Plagiarism Detector'[29], which gives a 'plagiarized' score from 0–100%, with a higher score indicating that more plagiarism was detected; these analyses were run in December. We also evaluated the abstracts using a paid similarity checker program, iThenticate[23], which outputs a 0–100% 'similarity index', with a higher score indicating more redundant with existing text. We also evaluated abstracts with an AI output detector using the 'GPT-2 Output Detector'[21,22], a RoBERTa-based sequence classifier, which gives abstracts a score ranging from 0.02 to 99.98% 'fake', with a higher score indicating the text was more likely to be generated by an AI algorithm.

We evaluated whether blinded human reviewers (study team members FMH, NSM, ECD, SR, members of our biomedical sciences laboratories used to reading scientific abstracts) could identify ChatGPT-generated abstracts. For every pair of reviewers, we used randomization via an electronic coin flip to decide whether an original or generated abstract would be provided for the first reviewer, with the opposite being given to the second reviewer. Each reviewer was given 25 abstracts to review, informed that there was a mixture of original and generated abstracts, asked to give a binary score of whether they thought the abstract was original or generated and invited to make free-text observations while reviewing. Reviewers were not shown any data or analysis until after their scoring of abstracts was completed.

We gave a binary yes/no score of whether the format of the generated abstract adhered to the journal's requirements by comparing it to the original article's headings and structure. We also compared the reported patient cohort sizes between the original and generated abstracts with a Pearson correlation of the logarithmic cohort sizes.

Statistics and visualization

Graphics and statistics were done in Python version 3.9 with *seaborn* version 0.11.2[30], *matplotlib* version 3.5.1[31], *sklearn* version 1.0.2[32], *scipy* version 1.7.3[33], and *statsannotations* version 0.4.4[34]. Group statistics are reported using median [interquartile range] and were compared using two-sided Mann Whitney Wilcoxon (MWW) tests, with $p < 0.05$ being the cutoff for statistical significance. Proportions were compared with Fisher's Exact tests. Correlation between the cohort sizes was done with Pearson's correlation.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The data used in the manuscript are available upon reasonable request to the corresponding author.

CODE AVAILABILITY

The code used in the manuscript are available upon reasonable request to the corresponding author.

Received: 1 January 2023; Accepted: 30 March 2023;
Published online: 26 April 2023

REFERENCES

1. OpenAI. ChatGPT: Optimizing language models for dialogue. *OpenAI* https://openai.com/blog/chatgpt/ (2022).
2. Shankland, S. ChatGPT: Why everyone is obsessed this mind-blowing AI chatbot. *CNET* https://www.cnet.com/tech/computing/chatgpt-why-everyone-is-obsessed-this-mind-blowing-ai-chatbot/ (2022).
3. Agomuoh, F. ChatGPT: how to use the viral AI chatbot that took the world by storm. *Digital Trends* https://www.digitaltrends.com/computing/how-to-use-openai-chatgpt-text-generation-chatbot/ (2022).
4. Hern, A. AI bot ChatGPT stuns academics with essay-writing skills and usability. *The Guardian* (2022).
5. Haque, M. U., Dharmadasa, I., Sworna, Z. T., Rajapakse, R. N. & Ahmad, H. "I think this is the most disruptive technology": exploring sentiments of ChatGPT early adopters using Twitter data. https://arxiv.org/abs/2212.05856 (2022).
6. Stokel-Walker, C. AI bot ChatGPT writes smart essays-should professors worry? *Nature* https://doi.org/10.1038/d41586-022-04397-7 (2022).
7. Whitford, E. Here's how Forbes got the ChatGPT AI to write 2 college essays in 20 min *Forbes* https://www.forbes.com/sites/emmawhitford/2022/12/09/heres-how-forbes-got-the-chatgpt-ai-to-write-2-college-essays-in-20-minutes/?sh=2b5a552456ad (2022).
8. Yeadon, W., Inyang, O.-O., Mizouri, A., Peach, A. & Testrow, C. The death of the short-form Physics essay in the coming AI revolution. https://doi.org/10.48550/ARXIV.2212.11661 (2022).
9. Kung, T. H. et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2**, e0000198 (2023).
10. Susnjak, T. ChatGPT: the end of online exam integrity? https://arxiv.org/abs/2212.09292 (2022).
11. Much to discuss in AI ethics. *Nat. Mach. Intell*. **4**, 1055–1056 (2022).
12. Thorp, H. H. ChatGPT is fun, but not an author. *Science* **379**, 313 (2023).
13. Flanagin, A., Bibbins-Domingo, K., Berkwits, M. & Christiansen, S. L. Nonhuman "authors" and implications for the integrity of scientific publication and medical knowledge. *JAMA* https://doi.org/10.1001/jama.2023.1344 (2023).
14. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature* **613**, 612 (2023).
15. Korngiebel, D. M. & Mooney, S. D. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. *NPJ Digit. Med*. **4**, 93 (2021).
16. Clark, E. et al. All that's 'human' is not gold: Evaluating human evaluation of generated text. in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Vol. 1: *Long Papers*) 7282–7296 (Association for Computational Linguistics, 2021).
17. Kreps, S., Miles McCain, R. & Brundage, M. All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *J. Exp. Political Sci.* **9**, 104–117 (2022).
18. Briganti, G. & Le Moine, O. Artificial intelligence in medicine: today and tomorrow. *Front. Med. (Lausanne)* **7**, 27 (2020).
19. Masa. SciNote Manuscript Writer-using Artificial Intelligence. *SciNote* https://www.scinote.net/blog/scinote-can-write-draft-scientific-manuscript-using-artificial-intelligence/ (Masa, 2020).
20. Writefull. *Writefull* https://www.writefull.com/. Accessed December 2022.
21. Solaiman, I. et al. Release strategies and the social impacts of language models. https://arxiv.org/abs/1908.09203 (2019).

22. GPT-2 Output Detector. https://huggingface.co/openai-detector. Accessed December 2022.
23. Turnitin, I. by Plagiarism detection software. https://www.ithenticate.com/. Accessed December 2022.
24. Howard, F. M. et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat. Commun.* **12**, 4423 (2021).
25. Banerjee, I. et al. Reading race: AI recognises patient's racial identity in medical images. https://arxiv.org/abs/2107.10356 (2021).
26. Bishop, J. M. Artificial intelligence is stupid and causal reasoning will not fix it. *Front. Psychol.* **11**, 513474 (2020).
27. Tyrrell, J. How easy is it to fool AI content detectors? *TechHQ* https://techhq.com/2023/02/how-easy-is-it-to-fool-ai-content-detectors/ (2023).
28. Hosseini, M., Rasmussen, L. M. & Resnik, D. B. Using AI to write scholarly publications. *Account. Res.* 1–9 https://www.tandfonline.com/doi/full/10.1080/08989621.2023.2168535 (2023).
29. Arcade, C. Plagiarism checker. *Plagiarismdetector.net* https://plagiarismdetector.net/ (2010).
30. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
31. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
32. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
33. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
34. Charlier, F. et al. trevismd/statannotations: v0.5. *Zenodo* https://doi.org/10.5281/ZENODO.7213391 (2022).

ACKNOWLEDGEMENTS

C.A.G. is supported by NIH/NHLBI F32HL162377. F.M.H. is supported by ASCO/Conquer Cancer Foundation and Breast Cancer Research Foundation Young Investigator Award 2022YIA-6675470300 and NIH/NCI K12CA139160. S.R. is supported by the Burroughs Wellcome Fund Early Scientific Training to Prepare for Research Excellence Post-Graduation (BEST-PREP). Y.L. reports effort support from the National Institute of Health/NCATS U01TR003528, NLM R01LM013337. A.T.P. reports effort support from the National Institute of Health/National Cancer Institute NIH/NCI) U01-CA243075, National Institute of Health/National Institute of Dental and Craniofacial Research (NIH/NIDCR) R56-DE030958, grants from Cancer Research Foundation, grants from Stand Up to Cancer (SU2C) Fanconi Anemia Research Fund–Farrah Fawcett Foundation Head and Neck Cancer Research Team Grant, and the Horizon 2021-SC1-BHC I3LUNG grant.

AUTHOR CONTRIBUTIONS

C.A.G. and F.M.H. conceived and designed the study, with contributions from N.S.M., Y.L., A.T.P. C.A.G. acquired, generated, and analyzed the data. F.M.H., N.S.M., E.C.D., S.R. reviewed the abstracts. C.A.G. drafted the manuscript. All authors reviewed the data and revised the manuscript.

COMPETING INTERESTS

A.T.P. reports no competing interests for this work, and reports personal fees from Prelude Therapeutics Advisory Board, Elevar Advisory Board, AbbVie consulting, Ayala Advisory Board, and Privo Therapeutics, all outside of submitted work. The remaining authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-023-00819-6.

Correspondence and requests for materials should be addressed to Catherine A. Gao.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.