JACC: CARDIOVASCULAR IMAGING
© 2024 BY THE AMERICAN COLLEGE OF CARDIOLOGY FOUNDATION
PUBLISHED BY ELSEVIER

VOL. 17, NO. 11, 2024

Letters

RESEARCH LETTER

Evaluating Patient-Oriented Echocardiogram Reports Augmented by Artificial Intelligence

The immediate release of echocardiogram reports to patients, as mandated by the 21st Century Cures Act, may cause unnecessary worry or confusion until clinicians provide explanations. This study examines whether ChatGPT, generative AI that has shown effectiveness in generating echocardiogram reports,[1] can help clinicians efficiently explain echocardiogram reports to patients.

We used a HIPAA-compliant ChatGPT instance with GPT4 to rewrite echocardiogram reports in plain language. Five cardiologists (M.S., A.F., D.B., A.V., and R.R.) who were board-certified in echocardiography rated AI-generated rewrites of 100 transthoracic echocardiogram reports from NYU Langone Health, including 20 reviewed by all for inter-rater reliability. Additionally, 12 participants without clinical backgrounds evaluated a separate set of 40 rewrites of echocardiogram reports generated for the study, including 10 reviewed by all for inter-rater reliability. Echocardiographers and nonclinical participants rated rewrites alongside original reports on separate surveys containing 5-point Likert scales adapted with permission from prior studies.[2-4] Median scores were reported for rewrites assessed multiple times. This work met criteria of the NYU Grossman School of Medicine for quality improvement projects and adhered to the Revised SQUIRE guidelines.

The median patient age for echocardiogram reports was 66 years (Q1-Q3: 58-75 years). Left ventricular systolic dysfunction was present in 23% (Wald 95% CI: 16%-30%). The AI-generated rewrites (median length: 1,216 characters; Q1-Q3: 995-1,438 characters) were longer than conclusions from echocardiogram reports (median: 792 characters; Q1-Q3: 600-948 characters) and shorter than full echocardiogram reports (median: 2,150 characters; Q1-Q3: 1,822-2,448 characters).

Echocardiographers rated whether each AI-generated rewrite could be accepted without edits. They responded "strongly agree" for 29% (95% CI: 20%-38%), "agree" for 44% (95% CI: 34%-54%), "neutral" for 13% (95% CI: 6%-20%), "disagree" for 14% (95% CI: 7%-21%), and "strongly disagree" for none. They rated the accuracy of all rewrites either "all true" (84%; 95% CI: 77%-91%) or "mostly correct" (16%; 95% CI: 9%-23%) and rated none "about half correct," "mostly incorrect," or "all false." Regarding rewrites with incorrect statements, none were rated "potentially dangerous," 8% (95% CI: 3%-13%) "must be corrected but not dangerous," 4% (95% CI: 0%-8%) "indeterminate need for correction," 4% (95% CI: 0%-8%) "unlikely to need correction," and none "insignificant."

When echocardiographers assessed the relevance of each rewrite, 76% (95% CI: 68%-84%) contained "all of the important information," 15% (95% CI: 8%-22%) "most," 7% (95% CI: 2%-12%) "about half," 2% (95% CI: 0%-5%) "less than half," and 0 "none." Regarding rewrites with missing information, none were rated "potentially dangerous," 5% (95% CI: 1%-9%) "must be corrected but not dangerous," 1% (95% CI: 0%-3%) "indeterminate need for correction," 6% (95% CI: 1%-11%) "unlikely to need correction," and 12% (95% CI: 6%-18%) "insignificant." Assessing whether quantitative information was represented appropriately in rewrites, echocardiographers rated 54% (95% CI: 44%-64%) "strongly agree," 36% (95% CI: 27%-45%) "agree," 2% (95% CI: 0%-5%) "neutral," 1% (95% CI: 0%-3%) "disagree," and none "strongly disagree."

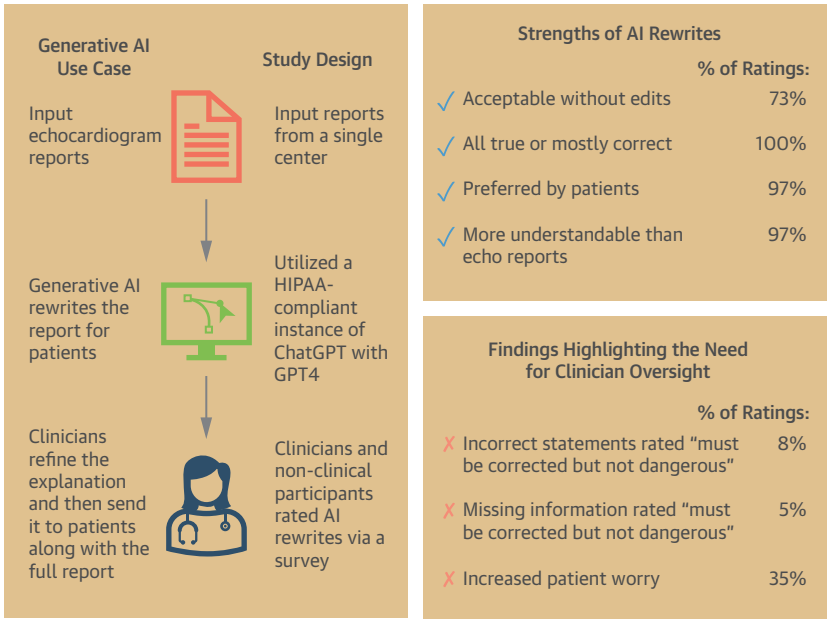
Nonclinical participants compared the understandability of each AI-generated rewrite to the original report and rated 70% (95% CI: 56%-85%) "much more," 27% (95% CI: 14%-31%) "a little more," and 3% (95% CI: 0%-7%) "equally" understandable. None were rated "a little less" or "much less" understandable. Regarding whether AI-generated rewrites would change their level of worry, participants rated 15% (95% CI: 4%-26%) "strongly reduce," 35% (95% CI: 20%-50%) "slightly reduce," 15% (95% CI: 4%-26%) "neutral," 30% (95% CI: 16%-44%) "slightly increase," and 5% (95% CI: 0%-12%) "strongly increase" level of worry. When participants rated whether they would prefer to have AI-generated rewrites in addition to original reports, 85% (95% CI: 74%-96%) "strongly prefer to have" and 12% (95% CI: 2%-23%) "slightly prefer to have" the rewrite, 3% (95% CI: 0%-7%) had "no preference," and none "slightly prefer not to have" or "strongly prefer not to have" the rewrite.

ISSN 1936-878X/$36.00

https://doi.org/10.1016/j.jcmg.2024.05.021

JACC: CARDIOVASCULAR IMAGING, VOL. 17, NO. 11, 2024
NOVEMBER 2024:1381-1383

FIGURE 1 Evaluating Echocardiogram Reports Rewritten in Nontechnical Language by Generative AI



AI = artificial intelligence.

Inter-rater reliability was poor among echocardiographers (intraclass correlation coefficient: 0.35; 95% CI: 0.25-0.46) and fair among nonclinical participants (intraclass correlation coefficient: 0.48; 95% CI: 0.32-0.6650) using a 2-way random-effects model, single rater type, and consistency definition.

Overall, most AI-generated rewrites performed well according to echocardiographers, who deemed 73% suitable for direct patient communication without modification (Figure 1). AI hallucination, or presenting false or misleading information as fact, was detected in 1 instance, in which the rewrite stated that a pleural effusion was small when size was not originally specified.

Nonclinical participants found 97% of AI-generated rewrites more understandable than the original reports. Enhanced understanding from rewrites reduced worry for 50% of participants, slightly increased it for 30%, and strongly increased it for 5%. Notably, the 2 instances where rewrites strongly heightened worry both involved critical findings. Nearly all participants (97%) also preferred receiving the rewrites alongside the original reports.

This study was conducted using a single proprietary model at 1 center, and modifications and testing may be necessary before wider implementation. Although these results align with studies on patient-friendly imaging reports,[5] future research should evaluate multiple AI models, including health care-specific ones, and explore various prompting strategies and data integration, such as including medical history to enhance rewrites. Our survey questions lack validation. Additionally, inter-rater reliability was low; training may enhance this. The echocardiographers involved in the study, who are also coauthors, may have introduced bias. Further research should assess how explanations with reports affect clinician workload.

Jacob A. Martin, MD, MSCR
Theodore Hill, BA
Muhamed Saric, MD, PhD
Alan F. Vainrib, MD
Daniel Bamira, MD
Samuel Bernard, MD
Richard Ro, MD
Hao Zhang, MS
Jonathan S. Austrian, MD
Yindalon Aphinyanaphongs, MD, PhD
Vidya Koesmahargyo, MS

Mathew R. Williams, MD
Larry A. Chinitz, MD
Lior Jankelson, MD, PhD*

*Leon H. Charney Division of Cardiology
Department of Medicine
NYU Grossman School of Medicine
550 1st Avenue
New York, New York 10016, USA
E-mail: Lior.jankelson@nyulangone.org

Dr Williams has received funding for research from Abbott, Medtronic, Boston Scientific, and Edwards Lifesciences. Dr Chinitz has received speaker fees/honoraria from Abbott Medical, Medtronic, Biotronik, and Biosense Webster. All other authors have reported that they have no relationships relevant to the contents of this paper to disclose.
The authors attest they are in compliance with human studies committees and animal welfare regulations of the authors' institutions and Food and Drug Administration guidelines, including patient consent where appropriate. For more information, visit the Author Center.

REFERENCES

1. Chao C-J, Banerjee I, Arsanjani R, et al. EchoGPT: a large language model for echocardiography report summarization. *medRxiv*. Published online January 20, 2024. https://doi.org/10.1101/2024.01.18.24301503

2. Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of ai-generated medical responses: an evaluation of the Chat-GPT Model. *Res Sq*. Published online February 28, 2023. https://doi.org/10.21203/rs.3.rs-2566942/v1

3. Liu S, Wright AP, Patterson BL, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *J Am Med Inform Assoc*. 2023;30:1237–1245.

4. Fabbri AR, Kryściński W, McCann B, Xiong C, Socher R, Radev D. SummEval: re-evaluating summarization evaluation. *Trans Assoc Comput Linguist*. 2021;9:391–409.

5. Lopez R, Kemp J, Lounsbury O, Short R, Befera NT. User-reported experience of patient-friendly imaging reports: opportunities to improve patient-centered outcomes in radiology. *J Am Coll Radiol*. 2024;21:40–43.