JACC: ADVANCES VOL. 2, NO. 8, 2023

EDITORIAL COMMENT

The Importance of External Validation for Neural Network Models*

Shinichi Goto, MD, PhD,[a,b] Hideki Ozawa, MD, PhD[a]

Machine learning with neural networks has evolved as a powerful tool to build clinically useful tools from complex medical data.[1] Recent studies have shown that the technology is capable of detecting diseases and predicting prognosis beyond the ability of fully trained experts.[2-4] However, the neural network model usually comes with the cost of the black box nature: not being able to explain why they predict what they predict. The characteristic not only prevented the application of neural networks for educating the experts with novel features extracted by the models but sometimes resulted in the development of useless models using unwanted features. For example, a model trained to detect pneumonia from chest X-ray images showed apparently high discrimination but was eventually found to have significantly lower performance on external datasets.[5] The study showed that the model could accurately identify the institution and setting where the X-ray was obtained. Since those who had the X-ray taken in an inpatient setting with a portable scanner had a significantly higher prevalence of pneumonia, the model presumably used the difference in the settings as a feature to detect pneumonia. This could have been easily avoided with a model that uses human-picked features.

Given the broad utility and the potential to develop powerful models, ensuring the generalizability of neural networks is an area of active research.[6-8] To reduce the black box problem, multiple techniques have been developed to identify the features utilized by neural network models, including gradient-weighted class activation mapping and local interpretable model agnostic explanations. While these techniques are able to partially explain the features utilized by the model, it only provide information on "where" (eg, within the QRS complex of the electrocardiogram [ECG]) the feature is thus far. It does not provide information on "what" (eg, is it the amplitude or duration?) it is. Thus, in our opinion, these techniques alone cannot guarantee that unwanted features are not used.

Another way of showing the robustness of the model is to test it directly on datasets with various backgrounds. In this issue of *JACC: Advances*, Harmon et al[9] have beautifully shown that their model previously developed to detect cardiac amyloidosis from ECG generalized well to a prospective dataset obtained after the model's development at the same institution. The same institution could share the artifact used as an "unwanted feature" if it existed. Thus, to purely evaluate the robustness of the model to unseen data, an external dataset would have been a better choice. However, the prospective nature of the current study supports that the feature used by the model to detect amyloidosis was not diminished by the advancement of treatment over time. The authors further support this finding by formally analyzing model performances over different time periods.

The subgroup analysis on age, sex, race, and ECG abnormalities is another strength of the article by Harmon et al Even though the model performs well across the overall population, there could be a subpopulation where it performs poorly. There is also a

*Editorials published in *JACC: Advances* reflect the views of the authors and do not necessarily represent the views of *JACC: Advances* or the American College of Cardiology.

From the [a]Division of General Internal Medicine & Family Medicine, Department of General and Acute Medicine, Tokai University School of Medicine, Isehara, Japan; and the [b]Division of Cardiovascular Medicine, Department of Medicine, Brigham and Women's Hospital/Harvard Medical School, Boston, Massachusetts, USA.

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.

ISSN 2772-963X https://doi.org/10.1016/j.jacadv.2023.100610

possibility that the model is indirectly utilizing these obvious factors, in which case a complex neural network model is usually not needed. For example, it is reported that neural networks can detect the age and sex of the patient accurately from ECGs,[10] which could be internally used to detect amyloidosis. By performing a subgroup analysis in these subgroups, Harmon et al nicely show that their model performs robustly across populations and that none of these factors played a dominant role in the model to detect cardiac amyloidosis. They have also identified those with left ventricular hypertrophy and left bundle branch block patterns as a population in which the model performs worse. This information is valuable to make clinical decisions when the model is deployed.

Selecting the diseases for which to apply the model is also an essential part of developing clinically valuable models. In our opinion, cardiac amyloidosis is one of the best use cases for neural network models in the cardiology field for 3 reasons.[11] First, cardiac amyloidosis is a disease that causes progressive heart failure that leads to death but can be treated to prevent disease progression if promptly diagnosed. Cardiac amyloidosis has a specific treatment and, thus, it is extremely important to discriminate between other causes of heart failure. Second, cardiac amyloidosis is underdiagnosed in the current system. It has been reported that a patient, on average, requires 6 months and visits to 3 doctors before they are diagnosed due to the difficulty of suspecting the disease from nonspecific symptoms.[12] Raising suspicion from an inexpensive test could improve the situation. And finally, the diagnosis of cardiac amyloidosis can be confirmed by subsequent diagnostic tests such as cardiac magnetic resonance imaging and 1-13C-pyruvate-scintigraphy. While these modalities are expensive and cannot be performed on everyone with heart failure, the neural network model can serve as a method to improve the pretest probability, making these tests cost-effective.

In summary, we commend Harmon et al[9] for performing this prospective validation of their neural network model for detecting cardiac amyloidosis from ECGs with subgroup analysis showing the robustness of their model in various populations. However, some limitations need to be pointed out. The first and largest limitation of the current analysis is the lack of validation by external institutions. Data from the same institution usually shares the same artifact. If these artifacts exaggerated the model's performance, they could not be detected by the current validation study. Second, the population was predominantly White and lacked diversity. This resulted in a very large confidence interval for the non-White population. The prevalence and cause of cardiac amyloidosis are known to have racial differences, so validation in a population with different races is extremely important. This study shows a good starting point for performing a validation study when a neural network model is developed. A future study with external institutions with a wider variety of populations concerning race would be an essential next step in validation.

FUNDING SUPPORT AND AUTHOR DISCLOSURES

Dr Goto was supported by grants from Tokai University School of Medicine Project Research and Internal Medicine Project Research, SECOM Science and Technology Foundation, and AMED (23hma922012h0001 and 23ym0126813j0002). Dr Ozawa has reported that he has no relationships relevant to the contents of this paper to disclose.

ADDRESS FOR CORRESPONDENCE: Dr Shinichi Goto, Division of General Internal Medicine & Family Medicine, Department of General and Acute Medicine, Tokai University School of Medicine, 143 Shimokasuya, Isehara 259-1193, Japan. E-mail: sgoto2@tsc.u-tokai.ac.jp.

REFERENCES

1. Goto S, McGuire DK, Goto S. The future role of high-performance computing in cardiovascular medicine and science -impact of multi-dimensional data analysis. *J Atheroscler Thromb*. 2022;29:559-562.

2. Goto S, Goto S. Application of neural networks to 12-lead electrocardiography—current status and future directions. *Circ Rep*. 2019;1:481-486.

3. Deo RC. Machine learning in medicine. *Circulation*. 2015;132:1920-1930.

4. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat Rev Cardiol*. 2021;18:465-478.

5. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med*. 2018;15:e1002683.

6. Goto S, Solanki D, John JE, et al. Multinational federated learning approach to train ECG and echocardiogram models for hypertrophic cardiomyopathy detection. *Circulation*. 2022;146:755-769.

7. Yagi R, Goto S, Katsumata Y, MacRae CA, Deo RC. Importance of external validation and subgroup analysis of artificial intelligence in the detection of low ejection fraction from

electrocardiograms. *Eur Heart J Digit Health*. 2022;3:654–657.

8. Hsu W, Hippe DS, Nakhaei N, et al. External validation of an ensemble model for automated mammography interpretation by artificial intelligence. *JAMA Netw Open*. 2022;5:e2242343.

9. Harmon DM, Mangold K, Suarez AB, et al. Post-development performance and validation of the artificial intelligence-enhanced electrocardiogram for detection of cardiac amyloidosis. *JACC: Adv*. 2023;2:100612.

10. Attia ZI, Friedman PA, Noseworthy PA, et al. Age and sex estimation using artificial intelligence from standard 12-lead ECGs. *Circ Arrhythm Electrophysiol*. 2019;12:e007284.

11. Goto S, Mahara K, Beussink-Nelson L, et al. Artificial intelligence-enabled fully automated detection of cardiac amyloidosis using electrocardiograms and echocardiograms. *Nat Commun*. 2021;12:2726.

12. Martinez-Naharro A, Baksi AJ, Hawkins PN, Fontana M. Diagnostic imaging of cardiac amyloidosis. *Nat Rev Cardiol*. 2020;17:413–426.

KEY WORDS artificial intelligence, machine learning, neural network, validation