ESC
European Society of Cardiology

European Heart Journal - Digital Health (2021) **2**, 561–567
doi:10.1093/ehjdh/ztab061

ORIGINAL ARTICLE

Spectrum bias in algorithms derived by artificial intelligence: a case study in detecting aortic stenosis using electrocardiograms

Andrew S. Tseng[†], **Michal Shelly-Cohen**[†], **Itzhak Z. Attia, Peter A. Noseworthy, Paul A. Friedman, Jae K. Oh and Francisco Lopez-Jimenez** *

Department of Cardiovascular Medicine, Mayo Clinic, 200 First Street Southwest, Rochester, MN 55905, USA

Received 25 March 2021; revised 3 May 2021; online publish-ahead-of-print 14 July 2021

Aims Spectrum bias can arise when a diagnostic test is derived from study populations with different disease spectra than the target population, resulting in poor generalizability. We used a real-world artificial intelligence (AI)-derived algorithm to detect severe aortic stenosis (AS) to experimentally assess the effect of spectrum bias on test performance.

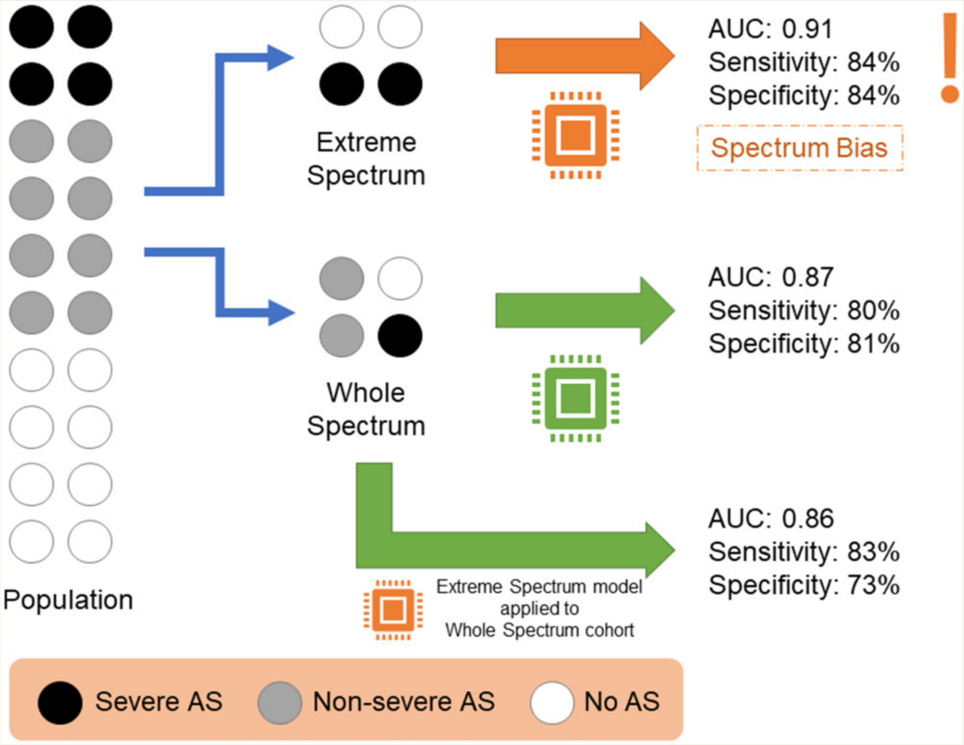
Methods and results All adult patients at the Mayo Clinic between 1 January 1989 and 30 September 2019 with transthoracic echocardiograms within 180 days after electrocardiogram (ECG) were identified. Two models were developed from two distinct patient cohorts: a whole-spectrum cohort comparing severe AS to any non-severe AS and an extreme-spectrum cohort comparing severe AS to no AS at all. Model performance was assessed. Overall, 258 607 patients had valid ECG and echocardiograms pairs. The area under the receiver operator curve was 0.87 and 0.91 for the whole-spectrum and extreme-spectrum models, respectively. Sensitivity and specificity for the whole-spectrum model was 80% and 81%, respectively, while for the extreme-spectrum model it was 84% and 84%, respectively. When applying the AI-ECG derived from the extreme-spectrum cohort to patients in the whole-spectrum cohort, the sensitivity, specificity, and area under the curve dropped to 83%, 73%, and 0.86, respectively.

Conclusion While the algorithm performed robustly in identifying severe AS, this study shows that limiting datasets to clearly positive or negative labels leads to overestimation of test performance when testing an AI algorithm in the setting of classifying severe AS using ECG data. While the effect of the bias may be modest in this example, clinicians should be aware of the existence of such a bias in AI-derived algorithms.

* Corresponding author. Tel: +1 507 284 8087, Fax: +1 507 266 7929, Email: lopez@mayo.edu

[†] The first two authors are shared first authorship and contributed equally to this project.

Graphical Abstract



Keywords Artificial intelligence • Aortic stenosis • Spectrum bias • Electrocardiogram

Introduction

With the advent of artificial intelligence (AI) and machine learning in clinical diagnostic testing, assessing their validity is paramount in applying this new technology to patient care. A major part of this assessment is the recognition of potential biases introduced throughout the model generation process, from training to validation to testing. It is important for the clinician to understand one of the major sources of bias, common in all such studies—the concept of spectrum bias.

The fundamental premise of spectrum bias in diagnostic testing is the use of an extreme spectrum of patients to derive and validate the test, even though a full spectrum exists and there is significant clinical heterogeneity within different subgroups of the disease. As such, a test may be derived from comparing patient cohorts from two extreme ends of the clinical spectrum (e.g. clearly normal versus severe or unequivocal disease), which on paper may improve test performance but results in poor generalizability in the real world where patients and the disease itself spans a full spectrum (e.g. mild, moderate or even indeterminate or equivocal disease severity).[1–3]

Examples abound of AI algorithms intended to diagnose clinical conditions where cases with clear cut disease were compared to normal controls. In one example on the use of AI in the detection of autism spectrum disorder from eye-tracking habits, researchers compared normal children and children with evident autism. The resultant model yielded a robust test accuracy of 89%.[4] Equally robust test performance for AI utilizing other behavioural habits between normal children and children with autism were noted in subsequent studies, with areas under the receiver operator curve of 0.89–0.93.[5] Yet, in these small studies, the spectrum of autism spectrum disease was not specifically evaluated, and it is possible that the test performance would not be as good in a general population of children.

Therefore, for clinicians to determine whether a particular AI-derived diagnostic tool is generalizable to their patients, the study methodology should be critically scrutinized. In this study, we provide an experimental case of an AI-derived algorithm using a convoluted neural network (CNN) using electrocardiograms (ECGs) to detect aortic stenosis (AS) manipulating the disease spectrum. The severity of AS is a spectrum ranging from normal, mild, moderate to severe, with clinical heterogeneity within each subgroup of severity. Using this experiment, we seek to test the effect of disease spectrum on test performance, to determine if spectrum bias would be also present when using an AI-derived diagnostic tool.

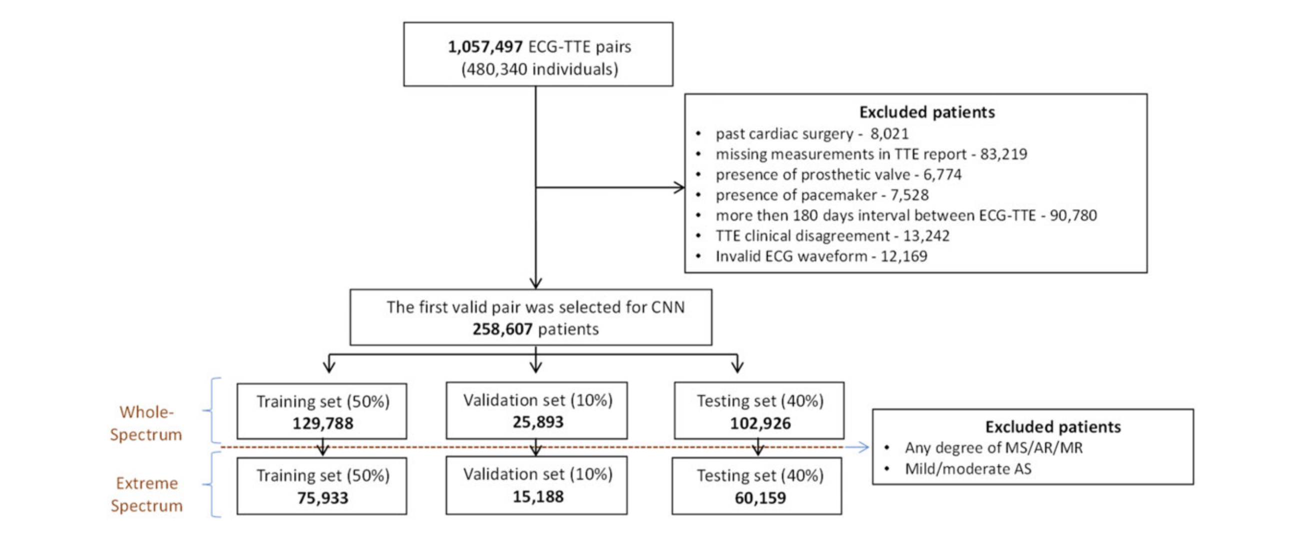


Figure 1 Flow diagram of the derivation of whole-spectrum and extreme-spectrum cohorts, including the number in the training, validation, and testing groups.

Methods

Cohort identification and study design

We identified all adult patients aged 18 years or older who had at least one transthoracic echocardiogram (TTE) and ECG performed at our institution between January 1989 and September 2019 using the Mayo Clinic Unified Data Platform that include tests from the Minnesota, Arizona, and Florida locations. The details are described in *Figure 1* and have been published previously.[6] Of the patients with TTE, we included only those with at least one of following AS measurements: aortic valve area, mean transaortic pressure gradient, peak transaortic velocity, or Doppler velocity index.[7,8] Of those, patients who had at least one digital, standard 12-lead ECG acquired within 180 days prior to their TTE exams were identified. When multiple TTEs and ECGs were available, we selected the first TTE available and the appropriate ECG-TTE pair to minimize the time interval between them. Patients with previous cardiac surgery, prosthetic valves, or pacemakers were excluded (*Figure 1*). Previous cardiac surgeries, particularly valve replacements, may alter the effects of AS on the ECG. Exclusion of pacemakers was based on the assumption that paced rhythms may not be informative as normal myocardial electrical activation is artificially perturbed.

The final patient cohort with valid ECG-TTE pair data was used for network creation, validation, and testing for the first ('whole-spectrum') model. We randomly assigned 50%, 10%, and 40% of our cohort to training, validation and testing the CNN, respectively (*Figure 1*). None of the patients was assigned to more than one group. For a second ('extreme-spectrum') model development, we excluded patients with any degree of aortic regurgitation, mitral regurgitation, mitral stenosis, mild AS, or moderate AS. Given the differences in dataset size between the whole-spectrum and extreme-spectrum cohorts, a secondary analysis was performed to balance the whole-spectrum cohort via random selection of patients to match the number of patients in the extreme-spectrum cohort. The model performance of this smaller balanced whole-spectrum cohort was then reassessed. The test performance of AI-ECG to identify severe AS patients for the whole-spectrum and extreme-spectrum models were compared to evaluate for the presence of spectrum bias. Furthermore, the decision threshold derived from the extreme-spectrum model was applied to the whole-spectrum cohort, simulating the test performance of a test derived from an extreme-spectrum disease spectrum on the whole-spectrum population.

To further validate that the performance effect uniquely arises from differences in disease spectrum and not simply from application of models to different hold-out cohorts, we further tested the whole-spectrum model on the extreme-spectrum cohort. If the performance degradation is only seen with the application of the extreme-spectrum model to the whole-spectrum cohort and not with the whole-spectrum model applied to the extreme-spectrum cohort, then there is likely a unique bias arising from disease spectrum.

The Mayo Clinic Institutional Review Board approved waiver of the requirement to obtain informed consent in accordance with 45 CFR 46.116 and waiver of Health Insurance Portability and Accountability Act authorization in accordance with applicable regulations.

Data sources and labelling

In order to characterize the study population and identify associated comorbidities at the time of ECG, the electronic health record was queried using standardized International Classification of Diseases, 9th and 10th Revision, billing codes for each diagnosis at any time prior to the index ECG date and within 30 days post-ECG. If a single code was found, the patient was considered to have that condition.

TTE data were extracted from the electronic health record and used to classify patients into two groups: TTE-positive AS were those with severe AS and TTE-negative AS were those with moderate mild or no AS by TTE for the whole-spectrum model, and no AS by TTE for extreme-spectrum model using AS severity grading guidelines (*Table 1*).[8,9] If a patient fulfilled any one of the following echocardiography parameters, the AS was classified as severe: peak velocity ≥4 m/s, mean gradient ≥40 mmHg, Doppler velocity index <0.25, or aortic valve area <1 cm^2. To make labelling more robust the physician impressions in the TTE report were also utilized. The physician impressions are standardized coded statements within our electronic database. Subjects with discrepancies between the final impressions and measurements were excluded.

Quantitative two-dimensional and Doppler echocardiography data were recorded using a Mayo Clinic–developed custom database. Mean

Table 1 Aortic stenosis measurements definitions

	Normal	Mild	Moderate	Severe
Aortic valve area, cm^2	2 or higher	(1.5–2.0)	[1.0–1.5]	Below 1.0
Peak transaortic velocity, m/s	2.5 or below	(2.5–3.0)	[3.0–4.0)	4 or higher
Transaortic mean pressure gradient, mmHg	10 or below	(10–20)	[20–40)	40 or higher
Doppler velocity index	0.5 or higher	(0.35–0.5)	[0.25–0.35]	Below 0.25

(Parenthesis) excludes the numbers shown and [bracket] includes them.

trans-aortic pressure gradient and peak velocity was acquired from all transducer positions to obtain the highest values.[8,10] Left ventricular outflow tract velocity and velocity-time integral were obtained and Doppler velocity index was calculated as a ratio between left ventricular outflow tract and aortic valve velocity-time integral.[7] Aortic valve area was calculated using the continuity equation.[7,8]

All ECGs were acquired as digital standard 10-s 12-lead ECGs using a Marquette ECG machine (GE Healthcare, WI, USA). The ECG waveform (raw data) was stored using the MUSE data management system for later retrieval.

Overview of AI model development

A CNN models using Keras framework with Tensorflow (Google; Mountain View, CA, USA) backend implemented in Python was developed.[11] Previously, we used this framework to create models to screen left ventricular contractile dysfunction and to estimate age as well as sex from standard 12-lead ECGs.[12,13] Each ECG was considered a matrix consisting of the following dimensions: 12 × 5000 (representing 12 leads for 10-s duration sampled at 500 Hz). The first dimension is spatial dimension and represents the different ECG leads and the second dimension is temporal. The 'Resample' function of the SCIPY python package was used to up-sample ECGs originally sampled in 250–500 Hz.[14] The CNN model is derived from a smaller version of DenseNet with 62 convolutional layers and 1 classification layer.[15] DenseNet uses densely connected convolutional blocks to concatenate the result of each convolutional output within the block in order to extract detailed features. We made minor modifications regarding zero padding to the original network to account for the difference in image and ECG matrix inputs.

We used the Adam optimizer for training with categorical cross-entropy as the loss function. Categorical cross-entropy was used even

Table 2 Patients characteristics and comorbidities

	Training set (*n* = 129 788)	Validation set (*n* = 25 893)	Testing set (*n* = 102 926)
Age, years (SD)	62.99 (16.3)	63.09 (16.3)	62.97 (16.3)
Age groups (%)			
<40	12 674 (9.8)	2508 (9.7)	10 094 (9.8)
40–49	12 978 (10.0)	2542 (9.8)	10 234 (9.9)
50–59	22 301 (17.2)	4466 (17.2)	17 909 (17.3)
60–69	31 231 (24.1)	6202 (24.0)	24 970 (24.2)
70–79	30 984 (23.9)	6242 (24.1)	24 077 (23.3)
80+	19 620 (15.1)	3929 (15.2)	15 642 (15.2)
Female sex (%)	61 514 (47.3)	12 288 (47.4)	48 988 (47.5)
Male sex (%)	68 274 (53.7)	13 605 (53.6)	53 938 (53.5)
AS measurement severity level (%)			
No AS	114 646 (88.3)	22960(88.7)	90 763 (88.1)
Mild AS	10 194 (7.9)	1991 (7.7)	8330 (8.1)
Moderate AS	1605 (1.2)	300 (1.5)	1225 (1.2)
Severe AS	3343 (2.6)	642 (2.5)	2608 (2.5)
Congestive heart failure (%)	23 399 (18.0)	4733 (18.3)	18 531 (18.0)
Peripheral vascular disease (%)	20 102 (15.5)	4178 (16.1)	16 134 (15.7)
Cerebrovascular disease (%)	14 787 (11.4)	3002 (11.6)	11 879 (11.5)
Renal disease (%)	15 641 (12.1)	3168 (12.2)	12 394 (12.0)
Chronic pulmonary disease (%)	26 312 (20.3)	5210 (20.1)	20 932 (20.3)
Connective tissue disease-rheumatic disease (%)	6273 (4.8)	1226 (4.7)	5103 (5.0)
Myocardial infarction (%)	12 097 (9.3)	2446 (9.4)	9843 (9.6)
Diabetes (%)	22 591 (17.4)	4563 (17.6)	18 186 (17.7)
Hypertension (%)	63 244 (48.7)	12 621 (48.7)	50 486 (49.1)

Any observed differences in comorbidities is a result of random chance.

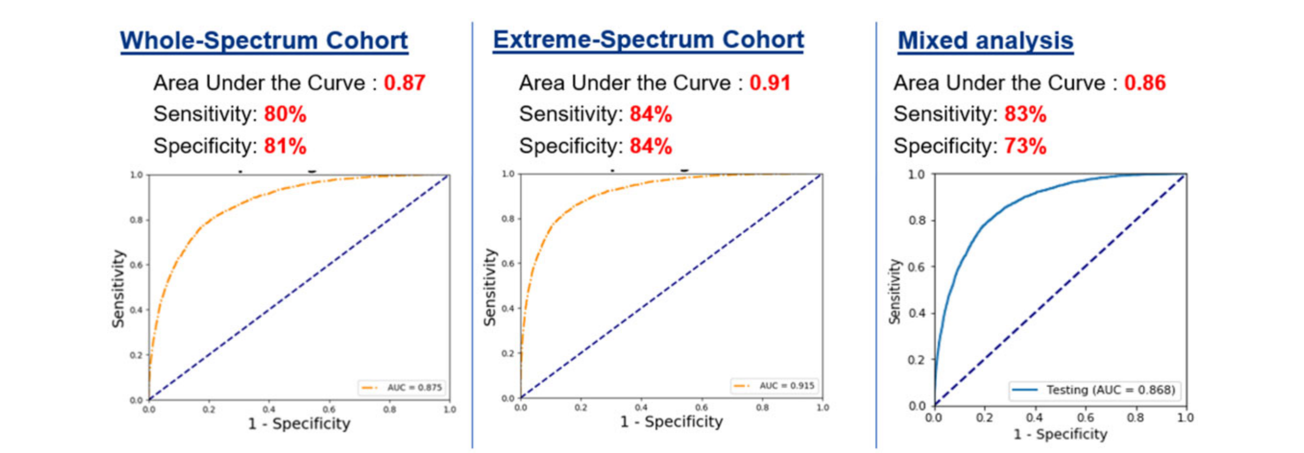


Figure 2 Receiver operating characteristic curves for three separate analyses with areas under the curve, sensitivity, and specificity, for the whole-spectrum cohort (left), the extreme-spectrum cohort (centre), and mixed analysis where the model derived from the extreme-spectrum was applied to patients from the general-spectrum cohort (right).

though it is a binary classifier due to the use of one hot encoding and having one output neuron for AI-ECG-positive AS and one for AI-ECG-negative AS. Hyper-parameters such as learning rate (1e-3) and batch size (64) were tuned using an internal validation set. We calculated an area under the curve (AUC) for the internal validation set after each epoch and the model with the highest AUC was used to test the holdout dataset.

To protect against biasing our estimate of the model performance, the training data was used exclusively for developing the model architecture. The threshold for classifying an ECG as either a positive or negative screen was determined using Youden index in the validation dataset. Once model training was completed, the final model performance was assessed using the testing data. We selected the CNN model architecture based on previous study on the same cohort where we aimed to detect moderate to severe AS.

Statistical analysis

Descriptive statistics were used to analyse the demographic and comorbidity data, Chi-squared test for categorical variables and the Student's *t*-test for continuous variables. Test performance analysis was derived from the testing data by constructing receiver operator curves. Test performance parameters (AUC, sensitivity, specificity, and accuracy) were derived with 95% confidence intervals using the large sample approximation of the DeLong method with optimization by the Sun and Xu method.[16] The optimal decision threshold via the Youden index was utilized as the probability cut-off for each derived model in the validation dataset.

Results

Baseline patient characteristics and comorbidities

Of 480 340 patients who had both TTE and ECG, 258 607 patients (54%) had valid ECG-TTE pairs. The derivation of the study cohorts is shown in *Figure 1*. The mean age was 63 ± 16.3 years with 122 790 (48%) women. The prevalence of TTE-confirmed severe AS was 2.6%. Of those with valid ECG-TTE pairs, 50% were used for training, 10% for validation, and 40% for testing. Patient characteristics and AS severity distribution were similar among the three cohorts (*Table 2*).

Test performance of the AI-ECG for detecting severe AS

The probability threshold for classifying an ECG as a TTE-positive AS screen in the validation data was determined to be 0.01635 and 0.03074 for the whole-spectrum and extreme-spectrum model, respectively, using the optimal decision threshold. Using these thresholds, the AUCs for identifying TTE-positive AS and TTE-negative AS subjects was 0.87 and 0.91 for the whole-spectrum and extreme-spectrum models, respectively, in both validation and testing groups (*Figure 2*). The secondary analysis to assess whole-spectrum model performance when the dataset size of the whole-spectrum cohort was balanced with the extreme-spectrum cohort resulted in the same AUC of 0.87 as the main analysis.

In the testing group, 2608 (2.5%) patients were labelled as AI-ECG-positive AS with a sensitivity and specificity for predicting echo-positive AS was 80% and 81% for whole-spectrum model and 84% and 84% for the extreme-spectrum model, respectively (*Figure 2*). This demonstrates that, while AI-ECG performed robustly in both models, the test performance was slightly reduced for the whole-spectrum model, though clinically this difference may not be significant.

When we applied the decision threshold for the extreme-spectrum model on the whole-spectrum cohort, the AUC results of 0.86 with sensitivity and specificity of 83% and 73%, respectively, lower than the AUC of 0.91 when using the extreme spectrum model in the corresponding extreme-spectrum cohort. This indicates a degradation in test performance when applying the extreme-spectrum model to the whole-spectrum cohort. This degradation in test performance was not seen when we applied the whole-spectrum model to the extreme-spectrum cohort, AUC 0.88, vs. 0.88 when using whole-spectrum model in the corresponding whole-spectrum cohort. The consistent reduction in test performance when AI-ECG is used on a cohort with all disease severities is suggestive of the presence of spectrum bias.

Discussion

We present the first study demonstrating the impact of spectrum bias in an AI-derived algorithm to detect severe AS using ECGs from a large cohort of patients at the Mayo Clinic. As the number of studies evaluating new diagnostic tools derived from AI algorithms continues to increase exponentially, it is important for clinicians to be able to critically evaluate these studies and determine their applicability in their practice. By using this example with AS we have illustrated the following: (i) the tangible impact of spectrum bias on test performance parameters, (ii) the importance of identifying key confounding variables, and (iii) the recognition of initial steps to reduce the impact of spectrum bias when interpreting studies.

Like in previous studies evaluating the test performance of various diagnostic tests, we have shown that spectrum bias may result in reporting results that are overestimate performance. Indeed, the AUC for the extreme-spectrum model was 0.91 with a sensitivity and a specificity of 84% and 84%, respectively. Because the learning cohort compared patients with vastly different demographics and comorbidities on the extreme ends of the AS spectrum (normal versus severe), we presented the model with a much easier binary classification problem.

When we repeated the machine learning process and introduced different severities of AS (mild and moderate), the test performance decreased, demonstrating the effect of spectrum bias on test performance. This shows that the AI algorithm performed better when the spectrum of disease was confined to the extremes, where subjects would not fall into mild or intermediate manifestations of the disease to be detected. When including patients from the complete disease spectrum, the algorithm is less able to identify or distinguish severe AS.

Furthermore, it should be noted that even when we applied the extreme-spectrum model on the whole-spectrum cohort, the test performance was still robust, with an AUC results of 0.86 with sensitivity and specificity of 83% and 73%, respectively. Therefore, even though the test performance was not as robust as that developed from the extreme-spectrum cohort, we may still be able to clinically utilize models derived from biased spectra, recognizing that performance would not be as good as shown in the original validation setting.

Therefore, it is critical to consider a few factors when interpreting studies in which spectrum bias may impact test performance. Firstly, there must be scrutiny of the cohorts used to derive the machine learning algorithm. What are the demographic and comorbidity characteristics of the derivation population and comparator groups? Did the investigators report all key variables known to impact test outcome? Secondly, is there a true binary classification (e.g. disease present/absent, pregnant non-pregnant, etc.) or does a range exists with a clinical or arbitrary threshold to define the presence of the disease? In the present study, AS severity exists on a spectrum that includes normal, mild, moderate, and severe. This applies to any potential confounding variable that is non-binary including tests with intermediary results or continuous variables such as ejection fraction, coronary flow, etc. Next, if a spectrum exists for the condition of interest, how did the investigators account for it during analysis? Are we able to generalize the results from the study population in our own patients?

There are multiple strategies to account for spectrum bias.[17,18] We had shown one common method in this study where patients from all disease severities were included in the learning and testing cohorts. The benefit of such a strategy is that the AI system learns from a more representative sample of patients, is trained to identify more features that differentiate labels and may therefore be more generalizable to real-world situations where patient mix is heterogeneous and not extreme-spectrum to extreme presentations of disease severity.

There were limitations in our study. We used a real-world example to demonstrate the general concept of spectrum bias in AI algorithms. It is likely that our specific findings may differ based on different types of machine learning methods, input data formats (i.e. numerical, graphical, etc.), clinical conditions of interest, or patient population. Secondly, apart from the demonstrating the AI-ECG's ability to assist in diagnosis, we were not able to evaluate the clinical utility on outcomes for such a test in this present study. Thirdly, our control group may include patients with significant cardiac structural abnormalities (such as reduced ejection fraction or other significant valvulopathies not involving the aortic valve). It is possible that more stringent exclusion criteria for our control group might have accentuated the spectrum bias noted in the present study. Next, while we used a standardized approach to identify severe AS, it is possible that there may be a small subset of patients who do not meet the study criteria for severe AS who may have been excluded from this study. To reduce this limitation, we used an inclusive definition of severe AS and improved its robustness by confirming with physician final impressions. Lastly, we acknowledge that spectrum bias in other fields of AI or other potentially encountered scenarios has not been established. Nonetheless, we provided a key example of spectrum bias using an AI-ECG model to bring this important concept to the attention of clinicians and researchers in this field.

Conclusion

Spectrum bias may be an important limitation in studies involving diagnostic tests and has been shown for the first time in an AI-derived testing algorithm to classify severe AS from ECG data. It is critical that clinicians recognize potential spectrum bias when reviewing these studies to ensure appropriate interpretation of the results and applicability in their own patient population.

Conflict of interest: Mayo Clinic has licensed the underlying technology to EKO, a maker of digital stethoscopes with embedded ECG electrodes. Mayo Clinic may receive financial benefit from the use of this technology, but at no point will Mayo Clinic benefit financially from its use for the care of subjects at Mayo Clinic. P.A.F., F.L.-J., and I.Z.A. may also receive financial benefit from this agreement. A.S.T., M.S.-C., and J.K.O. have no conflicts of interest to disclose.

Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

Lead author biographies

Andrew Sean Tseng is a cardiology fellow at the Mayo Clinic. In addition to his medical training, he obtained a Master's in Public Health from the Harvard T.H. Chan School of Public Health. His research interests include population health, outcomes research, economic analyses, particularly focusing on the potential impact of artificial intelligence on clinical practice within cardiology and its subspecialties.

Michal Cohen Shelly is an electrical engineer in the Mayo Clinic cardiovascular artificial intelligence team. She previously worked as a developer in private industry dealing with big data, business intelligence, science and communications. She joined the AI team in the Mayo Clinic Division of Cardiovascular Diseases in September 2018 and is involved in several high impact research projects using ECG and tabular data to accelerate AI-ECG research. Her main fields of interest are integrating patient needs and engineering as well as developing and leading innovative approaches to improve patient care.

References

1. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;**299**:926–930.
2. Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Ann Intern Med* 1992;**117**:135–140.
3. Jelinek M. Spectrum bias: why generalists and specialists don't connect. *ACP J Club* 2008;**149**:2.
4. Liu W, Li M, Yi L. Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework. *Autism Res* 2016;**9**:888–898.
5. Song DY, Kim SY, Bong G, Kim JM, Yoo HJ. The use of artificial intelligence in screening and diagnosis of autism spectrum disorder: a literature review. *Soa Chongsonyon Chongsin Uihak* 2019;**30**:145–152.
6. Cohen-Shelly M, Attia ZI, Friedman PA, Ito S, Essayagh BA, Ko WY, Murphree DH, Michelena HI, Enriquez-Sarano M, Carter RE, Johnson PW. Electrocardiogram screening for aortic valve stenosis using artificial intelligence. *Eur Heart J* 2021:ehab153.
7. Oh JK, Taliercio CP, Holmes DR Jr, Reeder GS, Bailey KR, Seward JB, Tajik AJ.. Prediction of the severity of aortic stenosis by Doppler aortic valve area determination: prospective Doppler-catheterization correlation in 100 patients. *J Am Coll Cardiol* 1988;**11**:1227–1234.
8. Baumgartner H, Hung J, Bermejo J, Chambers JB, Edvardsen T, Goldstein S, Lancellotti P, LeFevre M, Miller F Jr, Otto CM.. Recommendations on the echocardiographic assessment of aortic valve stenosis: a focused update from the European Association of Cardiovascular Imaging and the American Society of Echocardiography. *J Am Soc Echocardiogr* 2017;**30**:372–392.
9. Nishimura RA, Otto CM, Bonow RO, Carabello BA, Erwin JP 3rd, Guyton RA, O'Gara PT, Ruiz CE, Skubas NJ, Sorajja P, Sundt TM.. 2014 AHA/ACC guideline for the management of patients with valvular heart disease: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *J Am Coll Cardiol* 2014;**63**:2438–2488.
10. Thaden JJ, Nkomo VT, Lee KJ, Oh JK. Doppler imaging in aortic stenosis: the importance of the nonapical imaging windows to determine severity in a contemporary cohort. *J Am Soc Echocardiogr* 2015;**28**:780–785.
11. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM.. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging* 2016;**35**:1285–1298.
12. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, Pellikka PA, Enriquez-Sarano M, Noseworthy PA, Munger TM, Asirvatham SJ, Scott CG, Carter RE, Friedman PA.. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med* 2019;**25**:70–74.
13. Attia ZI, Friedman PA, Noseworthy PA, Lopez-Jimenez F, Ladewig DJ, Satam G, Pellikka PA, Munger TM, Asirvatham SJ, Scott CG, Carter RE.. Age and sex estimation using artificial intelligence from standard 12-lead ECGs. *Circ Arrhythm Electrophysiol* 2019;**12**:e007284.
14. Goehring C, Perrier A, Morabia A. Spectrum bias: a quantitative and graphical analysis of the variability of medical diagnostic test performance. *Stat Med* 2004;**23**:125–135.
15. Huang G, Liu Z, Pleiss G, Van Der Maaten L, Weinberger K. Convolutional Networks with Dense Connectivity. *IEEE Trans Pattern Anal Mach Intell* 2019:1–12.
16. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;**44**:837–845.
17. Bachmann LM, ter Riet G, Weber WE, Kessels AG. Multivariable adjustments counteract spectrum and test review bias in accuracy studies. *J Clin Epidemiol* 2009;**62**:357–361.e2.
18. Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Ann Intern Med* 2002;**137**:598–602.