Automated Detection of Aortic Stenosis Using Machine Learning

Benjamin S. Wessler, MD, Zhe Huang, MS, Gary M. Long, Jr, MS, Stefano Pacifici, MD, Nishant Prashar, MD, Samuel Karmiy, MD, Roman A. Sandler, PhD, Joseph Z. Sokol, Daniel B. Sokol, MS, Monica M. Dehn, RDCS, Luisa Maslon, RDCS, Eileen Mai, RDCS, Ayan R. Patel, MD, and Michael C. Hughes, PhD, *Boston, Medford, and Dorchester, Massachusetts; and Los Angeles, California*

Background: Aortic stenosis (AS) is a degenerative valve condition that is underdiagnosed and undertreated. Detection of AS using limited two-dimensional echocardiography could enable screening and improve appropriate referral and treatment of this condition. The aim of this study was to develop methods for automated detection of AS from limited imaging data sets.

Methods: Convolutional neural networks were trained, validated, and tested using limited two-dimensional transthoracic echocardiographic data sets. Networks were developed to accomplish two sequential tasks: (1) view identification and (2) study-level grade of AS. Balanced accuracy and area under the receiver operator curve (AUROC) were the performance metrics used.

Results: Annotated images from 577 patients were included. Neural networks were trained on data from 338 patients (average n = 10,253 labeled images), validated on 119 patients (average n = 3,505 labeled images), and performance was assessed on a test set of 120 patients (average n = 3,511 labeled images). Fully automated screening for AS was achieved with an AUROC of 0.96. Networks can distinguish no significant (no, mild, mild to moderate) AS from significant (moderate or severe) AS with an AUROC of 0.86 and between early (mild or mild to moderate AS) and significant (moderate or severe) AS with an AUROC of 0.75. External validation of these networks in a cohort of 8,502 outpatient transthoracic echocardiograms showed that screening for AS can be achieved using parasternal long-axis imaging only with an AUROC of 0.91.

Conclusion: Fully automated detection of AS using limited two-dimensional data sets is achievable using modern neural networks. These methods lay the groundwork for a novel method for screening for AS. (J Am Soc Echocardiogr 2023;■:■-■.)

Keywords: Aortic stenosis, Screening, Machine learning, Echocardiography, Transthoracic Echocardiography

Aortic stenosis (AS) is an enormous public health problem that affects more than 12.6 million adults and worldwide causes an estimated 102,700 deaths annually.[1] Recently, there has been interest in earlier identification of AS and evidence that many patients may not be appropriately treated.[2,3] These observations motivate the study of novel methods to identify AS. In this study, we evaluated whether machine learning (ML) methods can accurately identify AS using limited two-dimensional (2D) imaging data sets that are well suited for disease screening.

Little is known about how to improve the identification and treatment of AS. Using a population-based comprehensive transthoracic echocardiographic (TTE) screening approach would be prohibitively expensive. Automated interpretation of limited echocardiographic data sets is an attractive alternative approach to disease detection, especially with the rise of point-of-care ultrasound devices. Barriers to automating AS detection relate to the complex nature of this diagnosis, the need to integrate information across multiple images for any given study, and data sets that are not routinely annotated as part of routine clinical care.

From the CardioVascular Center, Tufts Medical Center, Boston, Massachusetts (B.S.W., M.M.D., L.M., E.M., A.R.P.); the Department of Computer Science, Tufts University, Medford, Massachusetts (Z.H., M.C.H.); CVAI Solutions, Dorchester, Massachusetts (G.L.); the Department of Medicine, Tufts Medical Center, Boston, Massachusetts (S.P., N.P., S.K.); and iCardio.ai, Los Angeles, California (R.A.S., J.S., D.B.S.).

This work is supported by National Center for Advancing Translational Sciences grant UL1TR002544. Dr Wessler is supported by National Institutes of Health grant K23AG055667.

This work was supported by the National Institutes of Health Tufts CTSI (NIH CTSA UL1TR002544). Dr Wessler received funding from the National Institutes of Health (K23 AG055667).

Dr Wessler has done consulting work with iCardio.ai and US2.ai unrelated to the present work and is a cofounder of CVAI Solutions. CVAI Solutions created the software for the deidentification procedures but currently has no related commercial pursuits.

Reprint requests: Benjamin S. Wessler, MD, Tufts Cardiovascular Center, Predictive Analytics and Comparative Effectiveness Center, Institute of Clinical Research and Health Policy Studies, Tufts Medical Center, 800 Washington Street, Box 63, Boston, MA 02111 (E-mail: *bwessler@tuftsmedicalcenter.org*).

0894-7317/$36.00

https://doi.org/10.1016/j.echo.2023.01.006

Abbreviations

2D = Two-dimensional

A4C = Apical four-chamber

AoV = Aortic valve

AS = Aortic stenosis

AUROC = Area under the receiver operating characteristic curve

ML = Machine learning

PLAX = Parasternal long-axis

PSAX = Parasternal short-axis

TMED-2 = Tufts Medical Echocardiogram Dataset, version 2

TTE = Transthoracic echocardiographic

Classically, accurate grading of AS relies on integration of numerous structural and hemodynamic parameters from across multiple imaging planes.[4] From the perspective of disease screening, certain features of AS, such as valve thickness and calcium burden, are readily apparent on 2D images. Although deep learning methods can now surpass humans in certain medical image classification tasks,[5,6] common classifier designs take only individual images as input, and applications in echocardiography have so far focused only on viewpoint identification, image segmentation, and assessments of ventricular function and myocardial diseases.[7-10] ML approaches to AS so far are limited to using echocardiography reports (thereby requiring expert image interpretation to work)[11,12] or are limited to very small numbers[13] or focus only on severe disease.[14] None have focused on assessing the continuum of AS severity using limited images with the goal of establishing tools suitable for automated disease screening. In this study we developed methods that can produce a coherent single diagnosis (severity of AS) from limited 2D data sets.

METHODS

Echocardiograms

This work was approved by the Tufts Medical Center institutional review board. The echocardiograms originate from TTE examinations performed between 2011 and 2020 at a high-volume tertiary care center (Tufts Medical Center). The echocardiograms were acquired as part of routine clinical care. The CardioVascular Imaging Center is Intersocietal Accreditation Commission accredited and is equipped with ultrasound units from major vendors (Philips, Toshiba, and Siemens). By using standardized Digital Imaging and Communications in Medicine images, these methods are intended to be vendor independent. Echocardiograms were included on the basis of the presence or absence of AS. Images were not selected for inclusion on the basis of image quality. Patients with prior aortic valve (AoV) replacements were excluded. Other cardiac findings, including other concomitant structural heart disease and rhythm abnormalities (i.e., atrial fibrillation), were not excluded.

Image Acquisition and Preprocessing

Images were acquired by trained sonographers with methods consistent with current American Society of Echocardiography guidelines.[15] For this study, we used metadata to identify and discard all spectral Doppler, color-flow Doppler, and M-mode recordings, keeping only 2D cardiac TTE images. To minimize computation time and enhance transportability and to position these networks for use in future screening environments, the first frame of each parasternal long-axis (PLAX) or parasternal short-axis (PSAX) AoV-level loop was automatically selected for use in the prediction models. If there were multiple PLAX or PSAX AoV-level acquisitions in a study (as is often the case), predictions used the first frame from each acquisition to arrive at study-level AS prediction. All images were standardized to 112 × 112 pixel resolution. Consistent with routine clinical care, there are no view or diagnostic label annotations available for images when they are collected.

Deidentification

Leveraging known region locations that are encoded within the Digital Imaging and Communications in Medicine storage format, propriety software was created to automatically identify the image burn region that contains protected health information. By excluding these imaging regions from the data copy, images were reliably deidentified. A 10% sample of the included deidentified images was manually reviewed to confirm that no protected health information was included.

Limited View Labels

The study setup followed the cognitive steps involved in diagnosing AS by echocardiography (Figure 1), specifically view recognition followed by view interpretation. We collected expert annotations of a limited number of view types with two goals in mind: (1) to evaluate (and validate) automated view classification networks and (2) to prioritize views for use in subsequent AS diagnostic models. Labels were assigned to examples of the PLAX and PSAX AoV-level views. These views were purposely selected because they are standard views that can visualize the AoV and can be prioritized in a limited screening environment. For evaluation of the view classification tasks, apical two-chamber (A2C) and apical four-chamber (A4C) views were also labeled. An "other" supercategory label that covered other 2D views was also collected. Doppler imaging was not included in this study, because these acquisitions are not routinely collected during point-of-care ultrasound imaging studies and because Doppler image acquisition requires a high level of skill that is often available only in dedicated echocardiography laboratories.

An echocardiogram annotation tool was built to facilitate view annotation (Supplemental Figure 1). Annotators (board-certified echocardiographers or American Registry for Diagnostic Medical Sonography–credentialed sonographers) assigned labels to more than two examples of each imaging view for each of the 599 studies included in our labeled set. Agreement between labelers was assessed on a set of 50 echocardiograms that were labeled in duplicate (Supplemental Tables 1 and 2).

Diagnostic Labels

The presence or absence of AS and the grade of AS (if present) were assigned by a cardiologist with specialty training in echocardiography. AS classification was assigned during clinical care in standard fashion following an integrative approach, as recommended in current guidelines (i.e., integrating information across all available images of all view types for a given patient).[4] The reference grade of AS for this study was taken directly from the clinical imaging report. AS labels for these experiments are shown in Table 1. Echocardiograms representing the full spectrum of AS pathologies were purposely included. To focus this work on potential automated screening use cases, and with recognition that interreader agreement of disease severity is modest,[16] we

HIGHLIGHTS

- Automated detection of AS is a novel approach to diagnosis.
- ML methods were trained to detect AS from limited 2D echocardiographic images.
- Fully automated screening for AS using limited data sets is achievable.
- Release of a TTE database will encourage collaboration.

grouped standard severity levels into three diagnostic classes: "no AS," "early AS" (combining mild and mild to moderate), and "significant AS" (combining moderate and severe). In a screening environment (upstream of the traditional echocardiography laboratory), the primary clinical question is which individuals should be referred for comprehensive echocardiography and AS-related care.

Data Sets and Experimental Design

Our experiments focused on assessing performance on echocardiographic studies from never-before-seen patients. These experiments were done in a manner consistent with the Proposed Requirements for Cardiovascular Imaging-Related Machine Learning Evaluation checklist.[17] The checklist for this study is available upon request.

Our final data set consisted of 599 fully labeled TTE studies, each of which has a diagnosis label (no AS, early AS, or significant AS) as well as some images with view labels. We have released this data set to researchers worldwide as the Tufts Medical Echocardiogram Dataset, version 2 (TMED-2). Most patients contributed only one study, but multiple studies from a small number of patients (22 of 577) were included to improve data set size. Each patient's data were assigned to exactly one set to properly assess generalization across individuals. The labeled data were divided into training (60%), validation (20%), and test sets (20%). We ensured that the ratio of diagnostic classes was the same across training, validation, and test (~21% no AS, ~29% early AS, and ~50% significant AS). Data set composition by label is summarized in Table 2 (for diagnosis task) and Supplemental Table 3 (for view task).

Each ML method was allowed to fit parameters to the training set, select hyperparameters on the basis of performance on the validation set, and report results on the test set. To improve the reliability of our results, we repeated the process of training a model and evaluating its performance across three separate, independent random partitions of all data into training, validation, and test sets. We report average performance across these three test sets. In addition to using the full training and validation set (479 studies), we also considered two levels of reduction (165 and 56 studies, roughly 33% and 11% of the full size). The same full-size test sets (120 studies) were always used to compare final performance.

Deep Neural Nets for View and Diagnosis Classification

We trained two neural networks: one view classifier and one diagnosis classifier. Each used the same backbone neural network architecture: a wide residual network with 28 layers containing 5,931,683 parameters.[18] Each network takes one image as input and produces a predicted probability vector. We discuss how we aggregate predictions across images in the next paragraph. The view classifier is trained to produce a five-way probabilistic view classification (PLAX, PSAX at the AoV level, A4C, apical two-chamber, or other) given a single image. To train, we minimize five-class cross entropy summed over all view-labeled images in the labeled set. The diagnosis classifier is trained so the same network produces two separate outputs given a single image: the primary output is a three-way probabilistic vector indicating the diagnosis (no, early, or significant AS), and the auxiliary output is a five-way probabilistic vector indicating the view type. We use multitask training, in which the loss function is a sum of the three-class diagnosis cross entropy and five-class view cross entropy, summed over all view-labeled images. We found that this multitask training delivered better diagnosis performance. After multitask training, only the three-class diagnosis output is used (the separately trained view neural network is a better view classifier than the auxiliary output). Each model was trained via stochastic gradient descent until the validation balanced accuracy for its primary task did not improve for at least 30 epochs. The actual epochs needed varied from 150 to 1,000 depending on the method used.

Producing One Study-Level Diagnosis from Many Images

Given the trained image-to-view and image-to-diagnosis classifier networks described previously, our goal was to automate the assignment of a study-level AS severity diagnosis: one vector

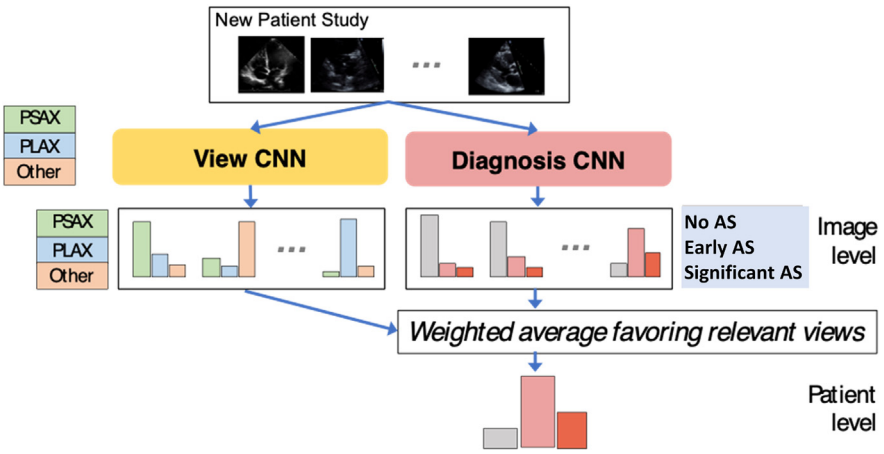


Figure 1 Approach to automated identification of AS. Convolutional neural networks (CNN) were trained and tested to identify view type and AS diagnostic category using limited 2D data sets.

Table 1 AS reference labels

Reference AS severity	Grading thresholds
Severe	Valve area < 1.0 cm^2, peak velocity > 4.0 m/sec, or mean gradient > 40 mm Hg
	Valve area < 1.0 cm^2, peak velocity < 4.0 m/sec or mean gradient < 40 mm Hg, and LVOT-derived stroke volume < 35 mL/m^2
Moderate	Valve area 1.0-1.5 cm^2, peak velocity 3.0-4.0 m/sec, or mean gradient 20-40 mm Hg
Mild	Valve area >1.5 cm^2, peak velocity 2.6-2.9 m/sec, or mean gradient <20 mm Hg

LVOT, Left ventricular outflow tract diameter.
AS severity was assigned using an integrative approach consistent with current American Society of Echocardiography guidelines.[4] The "severe" AS reference label includes both high-gradient and low-gradient subtypes. AS severity was pulled from the echocardiography report as assigned by the clinical reader. As part of routine care, an additional label of "mild to moderate" AS was assigned when hemodynamic profiles overlap the "mild" and "moderate" severity classes. This label was preserved for these experiments. Valve area represents the continuity equation–derived valve area.

summarizing the holistic interpretation of all images in a study. To accomplish this, we applied an approach we call prioritized view weighting,[19] which we developed on an earlier, smaller data set. The intuitive motivation is that diagnostic predictions made from images that show the AoV (PLAX or PSAX AoV-level views) should be considered stronger evidence than predictions of disease severity from other view types. Concretely, our prioritized view procedure obtains a study-level probabilistic prediction in three steps. First, using the image-to-diagnosis classifier to produce a three-class probability vector indicating AS severity for every image in the study. Second, use the image-to-view classifier to predict the probability of a relevant view (PLAX or PSAX) for every image. Finally, compute a weighted average over the three-class vectors from step 1, weighting each by the probability from step 2. We compare this prioritized view approach to an alternative simple average that treats diagnoses from all images equally without any weighting by the view classifier.

Performance Metrics

Throughout the evaluation of both view and diagnostic tasks, we use balanced accuracy as the performance metric of interest. Standard accuracy does not adequately assess performance when the data have imbalanced class distributions, as seen in both view and diagnostic tasks. Balanced accuracy is computed in two steps: compute the fraction of true members of each class that are correctly recognized, then average this fraction across all classes.

To further assess our method's utility as a screening tool, we use receiver operating curve analysis and report the area under the receiver operating characteristic curve (AUROC) for several potential use cases: (1) distinguishing no AS from any AS (early and significant), (2) distinguishing early AS from significant AS, and (3) distinguishing nonsignificant (no AS or early AS) from significant (moderate, moderate to severe, or severe) AS.

External Validations

External validation of the view classifier was done using the Stanford EchoNet Dynamic data set.[20] This data set contains 10,030 images of the A4C view type, gathered using completely different patient populations, clinical teams, and label assignments than our Tufts-focused data set. As all 10,030 images are A4C views, we report our view classifier's accuracy on this data set. We used all available A4C images in this data set, which are provided at 112 × 112 resolution (the same resolution we used for our images).

Two external validation studies of the AS diagnostic classifiers were done. The first was a temporal external validation performed on TTE examinations done at Tufts Medical Center from May to July 2022. These studies represent consecutive clinically indicated TTE examinations with AS classifications that were independently reviewed for this study (no AS, early AS, or significant AS, defined in an identical fashion to the model derivation tasks).

Next, we performed an external validation study on data provided by iCardio.ai. The data consisted of TTE examinations performed

Table 2 AS diagnosis label cohorts across training and test splits

	Number of studies				Number of labeled images			
	Total	None*	Early AS†	Significant AS‡	Total	None*	Early AS†	Significant AS‡
Training§ rowhead	360	76	103	181	10,253	999	1,316	7,938
Validation rowhead	119	25	34	60	3,505	344	402	2,759
Test rowhead	120	26	34	60	3,511	339	408	2,764

We show the number of echocardiographic studies assigned to training, validation, and test sets across all three possible AS severity levels for the fully labeled data set of 599 studies. Image counts represent the mean over three splits, as the exact number of images per study differs across splits. No split deviates by >12% from the reported mean here. Each patient's data were assigned to exactly one set to properly assess generalizability to new patients, while preserving similar proportions of each AS severity level across training and test.
*No AS.
†Mild and mild to moderate AS.
‡Moderate, moderate to severe, and severe AS.
§The training set also included an additional set with only view labels but no AS diagnosis label. This view-only set contained 705 studies representing 7,694 labeled images (see Supplement Table 3).

Table 3 Baseline characteristics of patients (n = 577)

Patient characteristic	Value
Age, y	74 (63-82)
Sex, female	43
Race	
Caucasian	85
Black	4
Latino	3
Other	8
Height, in	66 (63-69)
Weight, lb	174 (146-208)
BMI, kg/m^2	27.8 (24.2-32.1)
Systolic BP, mm Hg	129 (116-144)
Diastolic BP, mm Hg	72 (63-79)
Hypertension	80
Hyperlipidemia	68
Congestive heart failure	33
Diabetes	31
Prior myocardial infarction	13
Prior PCI	17
Prior CABG	13
Prior CVA	10
Current smoking	8

BP, Blood pressure; *CABG*, coronary artery bypass grafting; *CVA*, cerebrovascular accident; *PCI*, percutaneous coronary intervention. All baseline characteristics refer to the entire cohort of 577 patients. Data are expressed as median (interquartile range) or as percentage.

between 2018 and 2020 by an outpatient diagnostic imaging company. For this validation we had access to a single imaging view (PLAX) and the AS diagnostic label for the study. This validation was designed to test model performance on limited 2D acquisitions. The TTE studies for this cohort were performed for independent medical practices and clinics in more than 13 states in the United States, and data were acquired using ultrasound units from four major vendors (GE, Philips, Teratech, and Acuson). AS grade was assigned by a Core Cardiology Training Symposium level III echocardiographer.

Statistical Analysis

To evaluate the classification performance of each neural network, we report the balanced accuracy on the test set. To assess binary discrimination between two classes, we also report the area under the receiver operating characteristic curve. For each performance metric, we report a 95% CI computed using 5,000 bootstrapped samples of the test set. We average this reporting across three independent partitions or "splits" of the data into training and test.

RESULTS

The clinical characteristics of the patients included in this study are shown in Table 3. The primary labeled cohort included 577 patients. The median age was 74 years (interquartile range, 63-82 years). Forty-three percent of the patients were women. Eighty-six percent of the study population was Caucasian. The hemodynamic parameters of the echocardiograms are shown in Table 4. The median AoV peak velocity was 2.89 m/sec (interquartile range, 2.29-3.67 m/sec), the median peak gradient was 34.6 mm Hg (interquartile range, 21.0-54.0 mm Hg), and the average median gradient was 18.1 mm Hg (interquartile range, 11.9-30.6 mm Hg). The average left ventricular ejection fraction was 60% (interquartile range, 55%-65%).

Partitioning of the data set is shown in Table 2. The fully labeled set contained both diagnosis and view labels (599 studies representing 577 patients; 43,823 total images, of which 17,270 have view labels). After preprocessing, the median study in our data set contained 70 images (5th to 95th percentile range, 48-105 images; range, 15-181 images). The median number of images with view labels was 19 (5th to 95th percentile range, 4.9-71; range, 1-107). Fully labeled data were divided into training (60% [360 studies]), validation (20% [119 studies]), and test (20% [120 studies]) sets.

Table 4 Echocardiograms in the Tufts Medical Echocardiogram Dataset, version 2, data set

Screening task	All	No AS	Early AS		Significant AS	
AS grade		No AS	Mild AS	Mild to moderate AS	Moderate AS	Severe AS
n	599	127 (21.2)	144 (24.0)	27 (4.5)	132 (22.0)	169 (28.2)
Stroke volume, mL (n = 566)	59.6 (46-75.6)	54 (41-70.8)	61 (47-83.1)	50 (43.0-69.7)	66.0 (52.0-83.0)	59.0 (47.8-72.0)
LV ejection fraction, % (n = 599)	60 (55-65)	55 (45-60)	60 (55-65)	60 (55-65)	60 (55-65)	60 (55-65)
V_2 max, m/sec (n = 507)	2.89 (2.29-3.65)	1.73 (1.28-2.00)	2.45 (2.32-2.64)	2.89 (2.80-2.97)	3.26 (3.11-3.47)	4.32 (3.96-4.65)
AoV maximum gradient, mm Hg (n = 508)	34.4 (21-52.7)	12.1 (6.8-16.1)	24.1 (21.8-28.1)	34.1 (32.1-35.2)	42.6 (38.9-48.4)	74.4 (63.0-87.0)
AoV mean gradient, mm Hg (n = 491)	18.1 (11.9-30.0)	6.7 (4.1-8.5)	13.2 (11.6-15.1)	17.8 (15.8-19.4)	23.0 (20.2-26.2)	42.0 (35.0-50.0)

LV, Left ventricular; *V_2 max*, peak continuous-wave velocity.
Hemodynamic values were extracted from the medical record. Screening tasks correspond to the automated diagnostic tasks studied: (1) any AS (vs none), (2) early AS versus significant AS, and (3) no significant AS (vs significant AS). Not all values are available for every study; the number reporting each value is shown in the left-hand column. Data are expressed as number (percentage) or as median (IQR) unless otherwise specified. LV ejection fraction was assessed by integrating the biplane method of disk summation (modified Simpson's rule) with overall visual assessment.

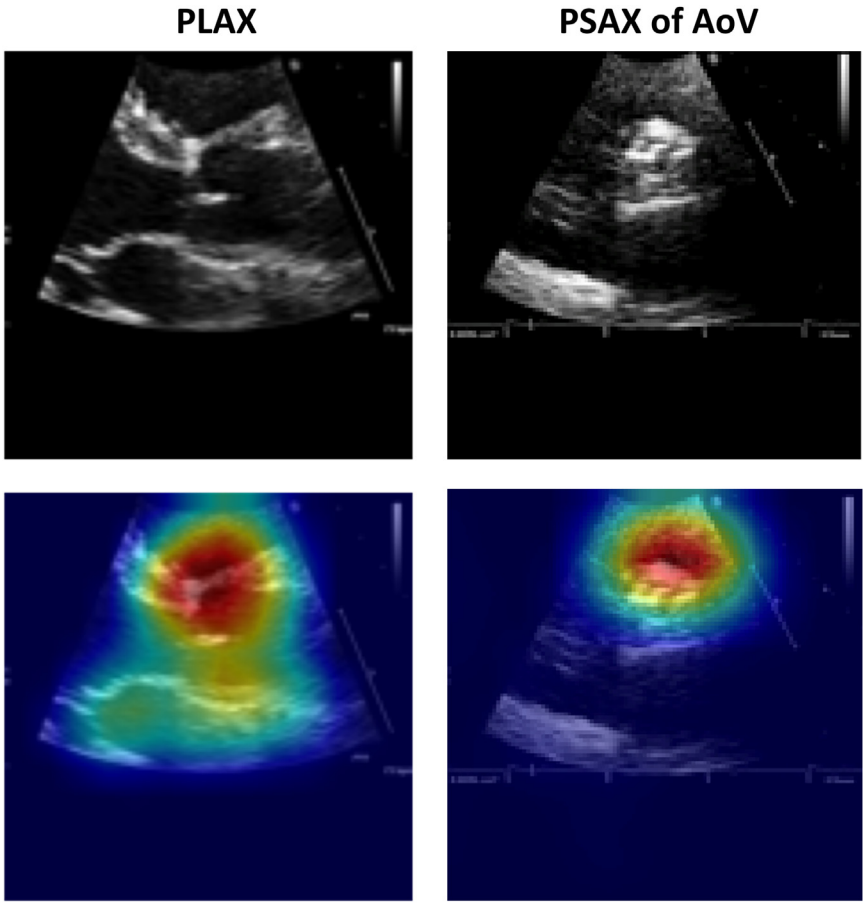


Figure 2 Grad-CAM visualizations of view predictions. Examples of PLAX *(left)* and PSAX AoV-level *(right)* views in our test set and their Grad-CAM visualizations. The original image is shown at the *top*, and the corresponding Grad-CAM visualization is shown *below* (original image with heat map overlay). The model correctly predicted the images to be PLAX and PSAX views, respectively, and correctly focused on the relevant region of the heart for making the predictions. The hotter the color, the more important the pixel in making the class discriminative decisions.

View Classification

Our view classifiers delivered 97% balanced accuracy on the test set when averaged over the three partitions of TMED-2. Balanced accuracy for the view task increased notably as the size of the available training and validation data increased from 90.3% with only 56 studies (95% bootstrap CI, 87.5%-90.4%) to all 97.0% with all 476 studies (95% bootstrap CI, 95.9%-97.5%; Supplemental Table 4). Using power-law curve fitting, which has been empirically successful at characterizing deep learning performance as data set sizes increase,[21] we project that labeled-set-only balanced accuracy could improve to 98.5% if 1,000 labeled studies were available for model training and validation (Supplemental Figure 2).

To sanity-check prediction quality, we used Grad-CAM[22] to generate visual explanation heat maps for our view classifier on select images from our test set (Figure 2). These visuals suggest that view predictions depend on relevant regions of the aortic root and AoV instead of irrelevant background data.

External Validation of View Classification

Accuracy at recognizing A4C views from the external EchoNet data set was 93.4% (95% bootstrap CI, 93.2%-93.8%) averaged over three

Table 5 Model discrimination for three AS screening tasks

Model	AS absent vs AS present	Early AS vs significant AS	Significant AS vs no significant AS
Simple average	0.86 (0.81-0.91)	0.73 (0.67-0.79)	0.84 (0.79-0.88)
Prioritized view	0.96 (0.93-0.97)	0.75 (0.68-0.81)	0.86 (0.82-0.90)

We report the AUROC for each task, averaged over three random training-validation-test splits of the data. The 95% bootstrap CI of this average is in parentheses. Two methods were used to aggregate image-level predictions to a study-level diagnosis: simple averaging or a weighted averaged that prioritizes specific views (PLAX or PSAX) that depict the AoV and are thus relevant for AS diagnosis.

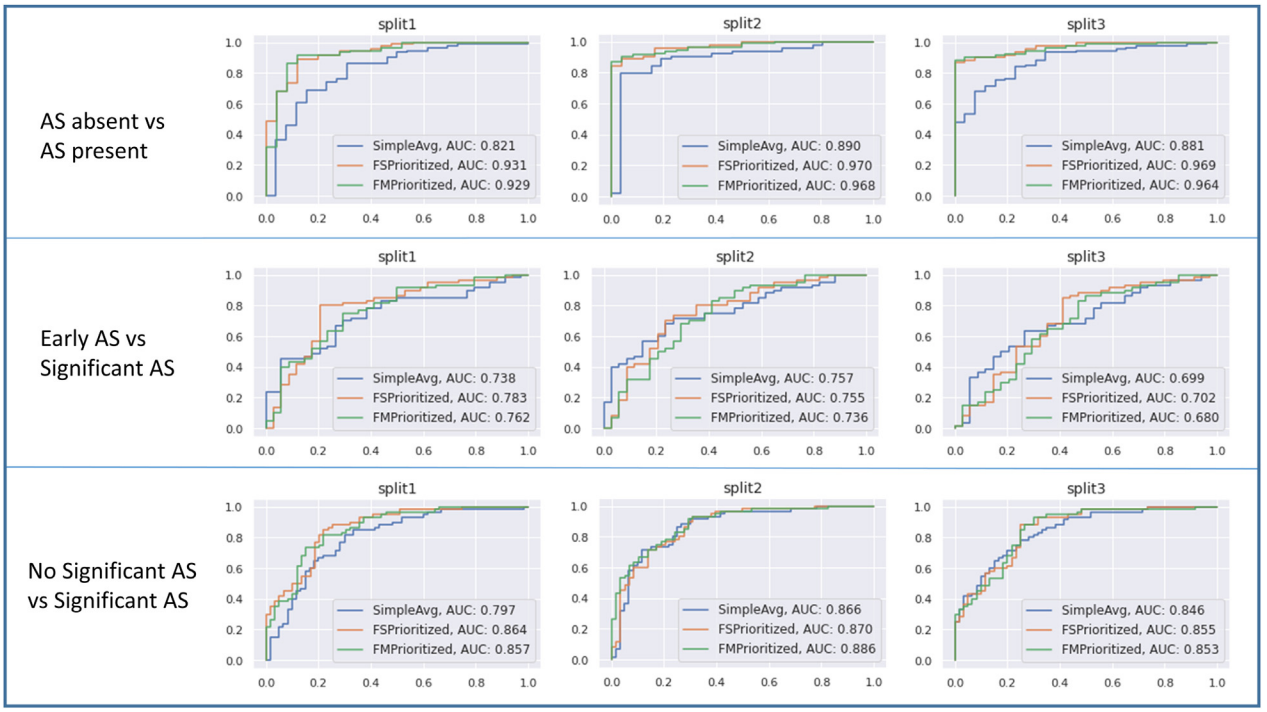


Figure 3 Diagnosis classification receiver operator curves. Each set of experiments was run with three random training-validation-test splits of the data (labeled split1, split2, and split3). The *top row* represents screening for AS: absent versus present (any severity). The *middle row* represents early AS (mild or mild to moderate) versus significant AS (moderate or severe). The *bottom row* represents nonsignificant AS (none, mild, or mild to moderate) versus significant AS (moderate or severe). Each *line* gives the performance of one prediction strategy for aggregating across all images in a study: prioritized view and simple average. Each *column* shows the results for one partition of the Tufts Medical Echocardiogram Dataset, version 2, data into training and test. *AUC*, Area under the receiver operating characteristic curve.

splits on the full labeled set (476 studies for training and validation, 120 for test). When the available labeled data were smaller, performance was naturally less accurate: 81.1% when developed on 165 studies and 66.1% when developed on 56 studies (Supplemental Table 5).

Diagnosis Performance Using Limited 2D Images Related to a Patient

Supplemental Figure 3 shows how study-level diagnosis classification improved with more labeled data across two strategies for averaging across all images to make a coherent study-level diagnosis (prioritized view vs simple average). On our full data set, the prioritized view strategy delivered 74.5% balanced accuracy for the three-way AS diagnosis task compared with 34.9% for simple average (Supplemental Table 6). Note that random guessing baseline would achieve 33% accuracy.

On the largest training split, the multitask training for our diagnosis classifier took about 14 hours on an Nvidia RTX6000 graphics processing unit. Using an already trained network, it takes approximately 0.4 sec to obtain an AS diagnostic prediction for a typical study.

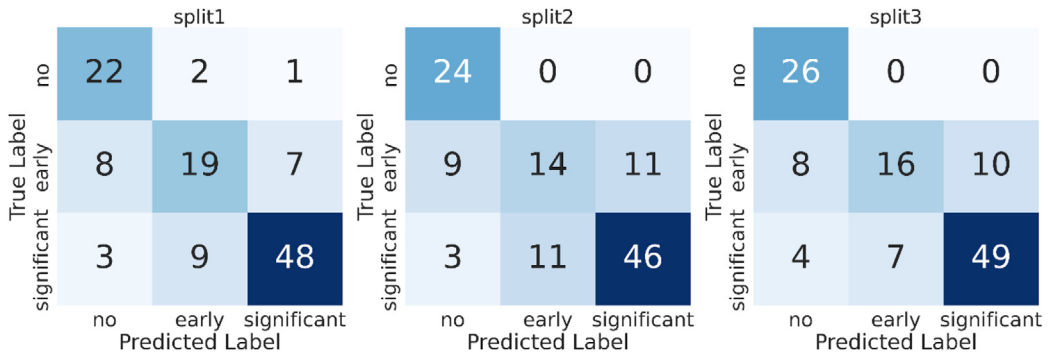
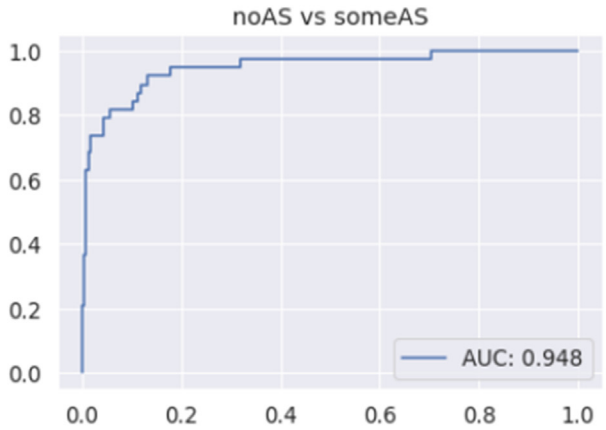


Figure 4 Confusion matrices for AS severity classification. Each set of experiments was run with three random training-validation-test splits of the data (labeled split1, split2, and split3). We report the test set confusion matrix from classifiers trained on each of the three training and test splits of our Tufts Medical Echocardiogram Dataset, version 2, data set.



Statistic	Point estimate	95% CI
Prevalence	14.5%	
Sensitivity	94.7%	82.5% to 99.4%
Specificity	78.7%	72.7% to 83.8%
PPV	42.9%	36.6% to 49.4%
NPV	98.9%	95.8% to 99.7%

Figure 5 Temporal external validation of the fully automated network for AS identification. The area under the receiver operating characteristic curve (AUC) is shown at *left* for 263 consecutive TTE examinations at Tufts Medical Center. AS diagnosis was independently reviewed for this study. *NPV*, Negative predictive value; *PPV*, positive predictive value.

Using Automatic Study-Level Diagnosis as a Preliminary Screening Tool for AS

Discriminatory performance for various clinical use cases are shown in Table 5 and Figure 3. Using limited 2D images, the AUROC was 0.96 for screening for any AS, 0.75 for identifying early (mild or mild to moderate) AS versus significant (moderate, moderate to severe, or severe) AS, and 0.86 for no significant AS (no AS and early AS) versus significant (moderate, moderate to severe, or severe) AS. Using these data, our methods demonstrate sensitivity of 88.3% and specificity of 88% for detecting AS. The confusion matrices for these predictions are shown Figure 4.

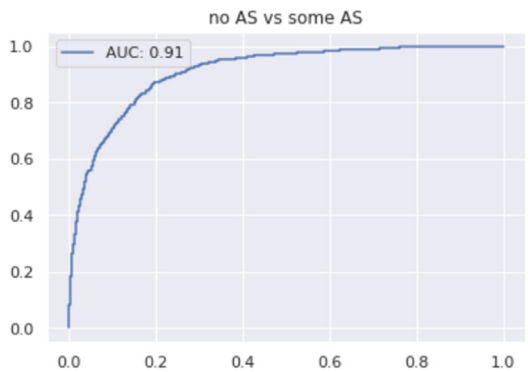
External Validation of AS Classification

In the temporal validation, the AS classifier was used to study 263 consecutive transthoracic echocardiograms acquired at Tufts Medical Center with independently verified AS grade as assigned by a board-certified echocardiographer. For the diagnostic screening task of identifying AS (all grades), the AUROC was 0.95. In this external validation cohort, the prevalence of AS was 14.5%. Sensitivity was 94.7% and specificity was 78.7%. Positive predictive value was 42.9% and negative predictive value was 98.9% (Figure 5). For the task of identifying significant AS (moderate and severe) versus no significant AS (no AS, mild AS, or mild to moderate AS), the AUROC was 0.95.

In the fully external validation using a single PLAX view, the AS classifier was used to screen for AS in 8,502 echocardiograms. For the screening task, the AUROC was 0.91. In this cohort, the prevalence of AS was 9.0%. Sensitivity was 89.3% and specificity was 76.1%. The positive predictive value was 27.0% and the negative predictive value was 98.6% (Figure 6).

DISCUSSION

Novel approaches to AS case identification are needed to improve treatment rates for this condition. Here we develop methods for fully automated detection of AS from limited TTE data sets. We show that automated detection of AS is possible using modern deep learning classifiers and that these networks are generalizable across different data sets. These tools can broadly characterize the presence or absence of AS and the severity of disease and are well suited for identifying patients who should be referred for comprehensive echocardiography. These results represent important steps toward establishing a novel approach to AS case identification.



Statistic	Point estimate	95% CI
Prevalence	9.0%	
Sensitivity	89.3%	86.9% to 91.4%
Specificity	76.1%	75.1% to 77.0%
PPV	27.0%	26.1% to 28.0%
NPV	98.6%	98.3% to 98.9%

Figure 6 Fully external validation on 8,502 echocardiograms from iCardio.ai. This validation study used only the PLAX view. AS diagnosis was assigned by fully independent echocardiographers. *AUC*, Area under the receiver operating characteristic curve; *NPV*, negative predictive value; *PPV*, positive predictive value.

These models are not designed to comprehensively phenotype AS, as can be done with complete transthoracic echocardiography. Instead, we view this work as a method to move case identification upstream of the echocardiography laboratory. With an estimated incidence rate of severe AS of 4.4% per year in the general population >65 years of age, it is clear that many patients go unrecognized.[23] The sensitivity and specificity of contemporary care with cardiac auscultation for detecting significant valve disease is only 44% and 69%, respectively.[24] Performance of auscultation is likely to be even lower for detecting more mild disease, in which murmurs are less intense. An automated screening program that uses limited 2D data sets—embedded within or upstream of hospital or clinic-based echocardiography laboratories—might improve case identification and referral. Although the discriminatory performance of these models appears excellent, the positive predictive value is modest. This is related in part to the relatively low prevalence of significant AS and should be viewed in the context of a very high negative predictive value (i.e., very few cases will be missed).

These tools could enable studies to address the profound treatment disparities for patients with severe AS[25,26] or interrogate emerging evidence that many patients with severe AS are not treated.[3,27,28] Automated screening could allow large-scale studies of the natural history of AS and also uncover potential biases in the care pathway of patients with this condition. Certainly, additional studies are needed to assess whether automation tools that enable effective screening and timely referral can improve outcomes for patients with AS. Automated detection of AS might also enable studies of early interventions to halt disease progression.[29] Classically, it has been challenging to study early stage disease, as early AS is asymptomatic. Fully automated interpretation of limited echocardiography may be worthwhile if effective treatments emerge or for enabling trial recruitment for treatment of earlier stage disease. The methods presented here do not use Doppler images and so are potentially suited to use with point-of-care ultrasound devices.

The modern networks studied here are attractive for the field of echocardiography because they can learn competitive models from small labeled data sets. These models used a single frame from the cine loops. Use of the time-varying feature sets almost certainly contains additional information, however this added information has to be balanced against the computational requirements needed to process more complex data sets. This network was designed to be scalable and require the least information necessary to be clinically valuable. Additionally, as demonstrated with the external validation study, these networks can ingest complete or partial studies and assign a diagnostic label. This flexibility positions these methods for use with limited acquisitions in screening environments. Here study-level diagnoses were achieved using a novel view-prioritized approach that uses a view classifier to identify views deemed relevant for the diagnostic task of interest (here AS). Diagnostic classifier predictions from these relevant views are then prioritized using a weighted average to predict a coherent study-level diagnostic label. We present validation studies of the sequential view and diagnostic tasks to emphasize the tiered approach used here that we believe can be applied to automate other complex imaging diagnoses.

The data used in these studies are released as part of TMED-2 (data and code are available at https://tmed.cs.tufts.edu). TMED-2 substantially increases the number of publicly released studies, increases resolution to 112 × 112 from 64 × 64, and increases available view types compared with our smaller earlier release.[19] This database covers a range of AS pathologies and will support the development of novel methods to automate screening for complex imaging diagnoses. The notable accuracy gains possible on external data with threefold increases in data set size illustrate the critical need for efforts to make labeled data sets available to researchers worldwide.

There were a few limitations to this work that must be recognized. The presented echocardiograms came from a single academic center, and diagnostic labels were assigned as part of routine clinical care. Nonwhite patients were underrepresented in this cohort, though the echocardiography-based imaging diagnosis of AS should not have any biologic differences on the basis of race. This study did not include outcome data or information from other imaging modalities to confirm disease severity. Although the number of labeled echocardiograms was modest, these networks are notable in that they can learn from small labeled sets. This is important for future model development where labels are expensive and time consuming. Although more complex low-flow, low-gradient subtypes may be misclassified, we minimize this risk by collapsing moderate and severe AS into a single "significant AS" category that should be referred for comprehensive study and care. This is by design and is important for future screening trials. Prior efforts that focus only on high-flow, high-gradient subtypes[14] would miss a significant number of cases that represent severe disease with lower flow profiles. With release of our code and images, we encourage additional external validations of our work. We expect that performance would improve with higher image resolutions, larger neural networks, or the use of all frames from cine loops rather than the first frame only; we kept resolutions modest (112 × 112 pixels) and used only one frame to achieve a tractable balance between accuracy and training time. On modern graphics processing units, each neural network we trained already requires dozens of hours on the largest version of our data set.

CONCLUSION

ML approaches optimized for echocardiography can successfully identify AS using limited 2D data sets. These methods lay the groundwork for fully automated screening for this disease and future study of interventions to improve outcomes.

SUPPLEMENTARY DATA

Supplementary data to this article can be found online at https://doi.org/10.1016/j.echo.2023.01.006.

REFERENCES

1. Yadgir S, Johnson CO, Aboyans V, et al. Global, regional, and national burden of calcific aortic valve and degenerative mitral valve diseases, 1990-2017. Circulation 2020;141:1670-80.
2. Lindman BR, Sukul D, Dweck MR, et al. Evaluating medical therapy for calcific aortic stenosis: JACC state-of-the-art review. J Am Coll Cardiol 2021;78:2354-76.
3. Li SX, Patel NK, Flannery LD, et al. Trends in utilization of aortic valve replacement for severe aortic stenosis. J Am Coll Cardiol 2022;79:864-77.
4. Baumgartner H, Hung J, Bermejo J, et al. 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;30:372-92.
5. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402-10.

6. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115-8.
7. Zhang J, Gajjala S, Agrawal P, et al. Fully automated echocardiogram interpretation in clinical practice. Circulation 2018;138:1623-35.
8. Ouyang D, He B, Ghorbani A, et al. Video-based AI for beat-to-beat assessment of cardiac function. Nature 2020;580:252-6.
9. Duffy G, Cheng PP, Yuan N, et al. High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning. JAMA Cardiol 2022;7:386-95.
10. Tromp J, Seekings PJ, Hung CL, et al. Automated interpretation of systolic and diastolic function on the echocardiogram: a multicohort study. Lancet Digit Heal 2022;4:e46-54.
11. Sengupta PP, Shrestha S, Kagiyama N, et al. A machine-learning framework to identify distinct phenotypes of aortic stenosis severity. JACC Cardiovasc Imaging 2021;14:1707-20.
12. Playford D, Bordin E, Mohamad R, et al. Enhanced diagnosis of severe aortic stenosis using artificial intelligence: a proof-of-concept study of 530,871 echocardiograms. JACC Cardiovasc Imaging 2020;13:1087-90.
13. Yang C, Ojha BD, Aranoff ND, et al. Classification of aortic stenosis using conventional machine learning and deep learning methods based on multi-dimensional cardio-mechanical signals. Sci Rep 2020;10:17521.
14. Dai W, Nazzari H, Namasivayam M, et al. Identifying aortic stenosis with a single parasternal long-Axis video using deep learning. J Am Soc Echocardiogr 2023;36:116-8.
15. Mitchell C, Rahko PS, Blauwet LA, et al. Guidelines for performing a comprehensive transthoracic echocardiographic examination in adults: recommendations from the American society of echocardiography. J Am Soc Echocardiogr 2019;32:1-64.
16. Haji K, Wong C, Neil C, et al. Multi reader assessment of accuracy and interobserver variability in aortic stenosis by echocardiography. Hear Lung Circ 2019;28:S258.
17. Sengupta PP, Shrestha S, Berthon B, et al. Proposed requirements for cardiovascular imaging-related machine learning evaluation (PRIME): a checklist. JACC Cardiovasc Imaging 2020;13:2017-35.
18. Zagoruyko S, Komodakis N. Wide residual networks. In: Procedings of the British Machine Vision Conference 2016; 2016;. pp. 87.1-87.12.
19. Huang Z, Long G, Wessler B, et al. A new semi-supervised learning benchmark for classifying view and diagnosing aortic stenosis from echocardiograms. Proc Mach Learn Healthc Conf, https://doi.org/10.48550/arXiv.2108.00080; 2021.
20. Ghorbani A, Ouyang D, Abid A, et al. Deep learning interpretation of echocardiograms. npj Digit Med 2020;3:10.
21. Hestness J, Narang S, Ardalani N, et al. Deep Learning Scaling is Predictable, Empirically. ArXiv, https://doi.org/10.48550/arXiv.1712.00409; 2017.
22. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE; 2017. pp. 618-26.
23. Durko AP, Osnabrugge RL, Van Mieghem NM, et al. Annual number of candidates for transcatheter aortic valve implantation per country: current estimates and future projections. Eur Heart J 2018;39:2635-42.
24. Gardezi SKM, Myerson SG, Chambers J, et al. Cardiac auscultation poorly predicts the presence of valvular heart disease in asymptomatic primary care patients. Heart 2018;104:1832-5.
25. Batchelor W, Anwaruddin S, Ross L, et al. Aortic valve stenosis treatment disparities in the underserved: JACC council perspectives. J Am Coll Cardiol 2019;74:2313-21.
26. Clark KA, Chouairi F, Kay B, et al. Trends in transcatheter and surgical aortic valve replacement in the United States, 2008-2018. Am Heart J 2022;243:87-91.
27. Tang L, Gössl M, Ahmed A, et al. Contemporary reasons and clinical outcomes for patients with severe, symptomatic aortic stenosis not undergoing aortic valve replacement. Circ Cardiovasc Interv 2018;11:1-12.
28. Brennan JM, Bryant A, Boero I, et al. Provider-level variability in the treatment of patients with severe symptomatic aortic valve stenosis. J Am Coll Cardiol 2019;73:1949.
29. Lindman BR, Merryman WD. Unloading the stenotic path to identifying medical therapy for calcific aortic valve disease. Circulation 2021;143:1455-7.